Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 28 No. 3s (2025)

Integrating Neural Network and SVM for Early Lifestyle Disease Diagnosis

Ms. Apeksha V. Sakhare[1],Dr. Prasad Lokulwar[2]

[1]Research Scholar, School of Engineering & Technology, GHRU Amravati University, Amravati.
apeksha.sakhare@raisoni.net

[2]Professor, Department of Computer Science and Engineering, G H Raisoni College of Engineering, Nagpur.
prasad.lokulwar@raisoni.net

Article History:

Received: *01-10-2024*

Revised: *22-11-2024*

Accepted: *01-12-2024*

Abstract:

The average life expectancy has increased by 15 to 20 years in recent decades due to major advancements in public health. Developments in public policy, economic growth, medical diagnoses, and treatment approaches have all contributed to these developments. However, the 21st century's fast urbanization and industrialization have changed food and lifestyle patterns, which has led to an increase in non-communicable diseases (NCDs). The fact that these diseases, which include obesity, type 2 diabetes, hypertension, and several types of cancer, currently cause 63% of deaths worldwide shows how pervasive they are.

Poor diet, inactivity, and sedentary work habits are behavioral variables that contribute to lifestyle-related illnesses (LRDs), a subgroup of noncommunicable diseases (NCDs) that include diabetes, stroke, and cardiovascular disease. Early detection and care are essential because these behaviors aggravate metabolic problems and other health issues. Despite being common, these illnesses can be difficult to diagnose because of their complexity and reliance on lifestyle and environmental factors. The requirement for a solid and dependable framework to accurately forecast and control LRDs is addressed in this study. It presents the NNSVM model, a hybrid method that combines feature selection, disease prediction, and practical suggestions for bettering one's lifestyle. The system uses lifestyle data and cutting-edge machine learning techniques to predict diseases like diabetes, depression, and hypertension in their early stages. This novel approach may strengthen efforts to prevent illness and assist people in changing their lifestyles in a way that will improve their health.

Keywords: selection of Features, Non communicable Diseases , Lifestyle Diseases, Bad Habits, Diseases Prediction, Neural Network, SVM.

Introduction

Non-communicable diseases (NCDs), sometimes referred to as lifestyle-related diseases, are responsible for over 63% of all deaths worldwide, underscoring their substantial influence on public health. Rather than infectious pathogens, these conditions—which include diabetes, hypertension, cardiovascular illnesses, and some forms of cancer—are predominantly caused by environmental and behavioral risk factors that may be changed. Some of the most prevalent causes of these chronic diseases are poor eating habits, sedentary lifestyles, smoking, excessive alcohol use, and extensive stress.

These disorders are frequently caused by unhealthy eating patterns, sedentary lifestyles, smoking, and binge drinking. The prevalence of these illnesses has sharply increased due to the fast industrialization and urbanization of nations like India, posing a serious public health concern. Chronic illnesses linked to lifestyle choices have become a major global health issue, placing a heavy burden on healthcare systems as well as on people's personal health.

Life Conventional diagnostic methods mostly depend on the identification of symptoms and clinical indicators, including blood pressure or glucose levels, which are frequently insufficient for early detection. The illness may have advanced to a point where therapy is more difficult and less successful by the time symptoms appear. This reactive strategy puts more strain on healthcare systems, raises healthcare expenditures, and leads to worse patient outcomes.
In the healthcare industry, machine learning (ML) has become a game-changing technology that has the potential to completely alter how medical data is examined, illnesses are identified, and treatments are administered. ML can process large and complicated datasets at a scale and speed that is significantly faster than human capacity by utilizing sophisticated algorithms. This capability enables healthcare systems to make more accurate and timely decisions by identifying patterns, trends, and correlations that could otherwise go overlooked.

However, current machine learning (ML)-based illness prediction models frequently ignore important lifestyle factors that have a substantial impact on the start and progression of disease in favor of concentrating only on clinical data. Furthermore, noisy, redundant, or insufficient data may cause these models to perform poorly.

This study fills these gaps by introducing a novel hybrid model called NNSVM, which combines the predictive powers of support vector machines and neural networks with the feature selection capabilities of random forest. By combining clinical data with lifestyle indicators like body mass index (BMI), smoking, alcohol use, stress levels, and patterns of physical activity, this model provides a comprehensive method for predicting diseases linked to a particular lifestyle. In order to enable prompt interventions, such as lifestyle changes and preventive measures, the suggested approach focuses on the early identification of chronic illnesses. The concept seeks to enhance overall health outcomes and slow the progression of diseases by proactively addressing these factors.

This method not only increases prediction accuracy but also emphasizes how crucial it is to use a variety of data sources in order to build reliable healthcare models. By encouraging proactive illness treatment and lowering long-term healthcare expenditures, the NNSVM model has the potential to revolutionize healthcare, as this study demonstrates through its creation, implementation, and evaluation [1].

The connection between dietary practices and lifestyle or sociodemographic characteristics has been the subject of numerous studies. Higher education or occupational prestige are frequently linked to healthier diets, while people with lower incomes, less prominent employment, or lower educational attainment are more prone to engage in unhealthy eating habits. Furthermore, poor eating habits are commonly linked to unhealthy activities including smoking, excessive alcohol use, and physical inactivity, especially in younger age groups. Furthermore, these habits frequently co-occur, resulting in unhealthy lifestyle clusters that are more common among people from lower socioeconomic backgrounds.

One of the most prevalent risk factors for cardiovascular issues in India is hypertension, or high blood pressure, which goes undetected or untreated in a significant section of the population. The chance of acquiring CVDs is further increased by diabetes and obesity, two conditions that are becoming more and more of a problem in India.

Government programs have been put in place to increase awareness, aid in early diagnosis, and promote better lifestyle choices. One such program is the National Programme for Prevention and Control of Cancer, Diabetes, Cardiovascular Diseases, and Stroke (NPCDCS). Notwithstanding these initiatives, sustained and targeted measures are necessary to address the rising incidence of non-communicable diseases (NCDs) and cardiovascular disease. In order to reduce the burden of chronic disorders and increase population well-being, community-driven initiatives that encourage healthy living, better healthcare facilities, and strong public health policies are essential.

One dichotomous variable that was included was the use of dietary supplements. Anthropometric data obtained during the physical examination were used to compute the waist circumference and BMI [2][3]. BMI people in younger age ranges. Furthermore, these habits frequently co-occur, resulting in unhealthy lifestyle clusters that are more common among people from lower socioeconomic backgrounds. However, as evidenced by research such as those employing latent class analysis, comparable patterns of clustered behaviors, such as mixed, unhealthy, and healthy habits, have been noted across many socioeconomic strata. The degree to which specific hazardous behaviors are similarly connected within low, middle, and high educated groups is yet unknown, despite the fact that such clusters are clearly present throughout a range of educational levels. The necessity for research into how these connections vary by socioeconomic characteristics, especially education, is highlighted by this knowledge gap [4].

1. RELEVENT WORK

This study says by 2025, 2.5 lakh instances of breast cancer are anticipated to be the most prevalent type in India. Diabetes is a disease that develops when blood glucose, often known as blood sugar, is extremely high. The primary energy source that you obtain from the food you eat is blood sugar. The hormone insulin is secreted by the pancreas and helps to operate metabolic processes by removing glucose from diet.[5]. There are 41 lakh diabetes-related deaths worldwide, according to the "International Diabetes Federation," and diabetes-related medical expenses total over 760 billion USD.[6] According to the Epidemiology of Diabetes, diabetes claims the lives of more than 10 lakh persons in India each year. Gavin Pinto, Radhika Desai, and Sunil Jangid's paper, "Understanding the lifestyle of people to identify the reasons for Diabetes using data mining," covered diabetes sub-classification as well as lowering the risk of diabetes disease with data mining approaches. On the dataset gathered through a Google Forms survey, the authors employed the Naïve Bayes and SVM classification algorithms. They reported that the accuracy of SVM was 65.93 and that of Naïve Bayes was 61.44.

The cardiac illnesses were thoroughly explained in the paper by M. Marimuthu, S. DeivaRani, and Gayatri. R. also used classification techniques including SVM, Decision Tree, Naïve Bayes, and K-Nearest Neighbors on the Framingham dataset from Kaggle. In order to predict the risk of heart disease, the scientists compared a number of machine learning methods. The KNN classification algorithm in this work has the highest recorded accuracy of 83.60% [7]. Richa Sharma and Dr. Kanak

Saxena address cardiovascular disease in the Purushottam-proposed work by utilizing Knowledge Extraction based on Evolutionary Learning, a Java programming technique for creating the development model for data mining problems. In this work, the maximum accuracy recorded is 86.7% [8][9]. Using the Nave Bayes classification method, M. Chinna Rao, K. Ramesh, and G. Subbalakshmi presented a decision support system for heart disease prediction. They covered the extraction of hidden information from heart disease datasets that can answer complicated questions [10]. A study by Amandeep Kaur and Jyothi Arora examined algorithms like KNN, SVM, ANN, and Decision Tree on the dataset of heart disease and plotted the accuracy graph [11]. Noreen Fatima highlighted the research of earlier models and advocated work on the cancer forecast using data mining and machine learning approaches that may accurately predict cancer on massive health records.

Shaik Subhani, Ch. Shravya, and K. Pravallika presented their work on breast cancer prediction using supervised machine learning algorithms on the dataset. They also used dimensionality reduction and principal component analysis (PCA) to examine the results and provide a well-organized explanation [12].The main topic of debate for Nikitha Rane and Jean Sunny's work on the categorization of cancer using machine learning ideas was the importance of early cancer detection in order to save many lives [13]. Using classification approaches, Dilip Singh Sisodia and Deepti Sisodia predicted diabetes with an accuracy of almost 76% on the Pima dataset.

Millions of people worldwide suffer from type 2 diabetes (T2D), which is defined by elevated blood glucose levels brought on by insulin resistance or insufficiency. Serious health issues like heart disease, stroke, nerve damage, renal disease, and visual loss can develop as type 2 diabetes worsens [14]. By slowing the evolution of type 2 diabetes, early detection and efficient treatment can lower the risk of complications and enhance patient outcomes. Prediction models based on artificial intelligence (AI) have been created to help medical practitioners identify the beginning and course of diabetes as well as other illnesses like seizures [15] and coronavirus disease 2019 (COVID-19) [16]. The models have the potential to transform diabetes care by facilitating early detection and intervention.

However, the scope of these studies is restricted to a detailed examination of the benefits and drawbacks of using AI algorithms to forecast the course of diabetes. The taxonomies pertaining to diabetes [17], [18], or algorithms [6] were the focus of the previous surveys. Though the overview of current issues was not thoroughly covered, the key results with intriguing future prospects on the evolution of diabetes were highlighted [19], [20]. Thus, using three different approaches—mathematical, machine learning (ML), and deep learning (DL)—this research will carefully investigate the proper methods for building prediction models.

Therapeutic inertia or delays in escalating T2D treatment could arise from improper T2D (Type 2 Diabetes) management, which could have expensive repercussions [21]. Similarly, it is anticipated that global direct health spending on T2D will rise from USD 760 billion in 2019 to USD 845 billion by 2030 [22]. As a result, very accurate diabetes progression prediction may aid in early identification and allow medical personnel to customize treatment regimens, thereby lowering treatment costs and improving patient outcomes [23]. With the development of AI technology, diabetes management and progression prevention are increasingly dependent on its prediction models. AI has the potential to

lower the present 8.8% global prevalence of diabetes by transforming its detection, prevention, and management [24].

An algorithm for machine learning is called Support Vector Machines (SVM). Many studies have presented SVM as a potent classification technique in recent years. It can be applied to regression [25,26] and is summarized in. According to earlier studies, SVM performs binary classification with a low error rate by using a high dimension space to find a hyperplane [27,28]. Using a function derived from the available training data to distinguish between the two classes is the challenge with SVM. Producing classifiers that perform well on other challenges is the goal. Two regions that make up the hyperplane function in SVM are separated by maximal input vectors.

Traditional models and methods of prediction include a number of risk factors and a variety of algorithmic metrics, including programs, datasets, and much more. Based on the results of group testing, patients are classified as high-risk or low-risk. However, these models are not useful in large industry sectors; they are only useful in therapeutic settings. Therefore, we have built the predictions system using the ideas of machine learning and supervised learning techniques in order to incorporate the illness forecasts in many health-related industries.

The substantial impact of lifestyle factors on the start and progression of non-communicable diseases (NCDs) is frequently overlooked by existing disease prediction models, which mostly depend on clinical symptom data. Since symptoms usually appear at advanced stages of illness progression, this symptom-based approach reduces the possibility of an early diagnosis. Furthermore, a lot of conventional models don't incorporate a variety of data sources, like lifestyle choices, which are important for determining the risk of disease.

The patient's symptoms were the only focus of earlier prediction techniques. A large number of the researchers concentrated on machine learning methods. Here, we substitute the patient's negative behaviours or lifestyles for the data on their symptoms.

In order to reduce death rates, lifestyle choices can aid in the early diagnosis of diseases. We can employ random forest feature extraction with neural networks and support vector machines to get over the limitations of the current approach. The model's performance can be assessed using performance evaluation metrics. Selecting several characteristics or features to predict lifestyle diseases is the primary problem here.

Numerous diseases that contribute significantly to world mortality are caused by environmental factors and human lifestyle choices, and diagnosing these illnesses can occasionally be challenging. We require a reliable, realistic, A unified model that can predict multiple LRDs, such as depression, diabetes, and hypertension, is crucial for holistic healthcare management, yet traditional frameworks mostly concentrate on single-disease prediction. The suggested framework uses a multi-output machine learning strategy that makes use of deep learning architectures like multi-task neural networks or ensemble techniques like Random Forest and Gradient Boosting. By recording the distinct patterns of each disease and sharing representations across them, these methods enable the simultaneous prediction of numerous disorders.

Furthermore, hybrid approaches that combine domain-specific knowledge with machine learning can improve the models' interpretability. For example, combining probabilistic models for estimating

illness risk with rule-based systems for early warning signals might give patients and healthcare professionals useful information.

2. PROPOSED METHODOLOGY

A mechanism for predicting models of non-communicable diseases is presented in this paper. Breiman introduced the random forest algorithm in 2001. A straightforward and effective machine learning technique, the approach has been widely utilized for data classification and result prediction in a variety of industries and fields, including biology, chemistry, geography, the Internet, and urban building , because of its notable performance. K decision trees are used by the random forest method to train and forecast data. There are two steps in its primary process:

Step 1: A decision tree is created for each training sample after K training samples are chosen at random from the input of the samples.

Step 2: Classify each decision tree node in a way that maximizes each tree's potential growth.

If the sample's feature dimension is M, a constant m much smaller than M is chosen. In order to train K decision trees, this study randomly picked m feature subsets from M features and determined which subset was best for each split. Lastly, a more accurate estimation of the prediction objective is obtained by integrating the prediction outcomes of several decision trees. The Random Forest algorithm's flow is as follows:

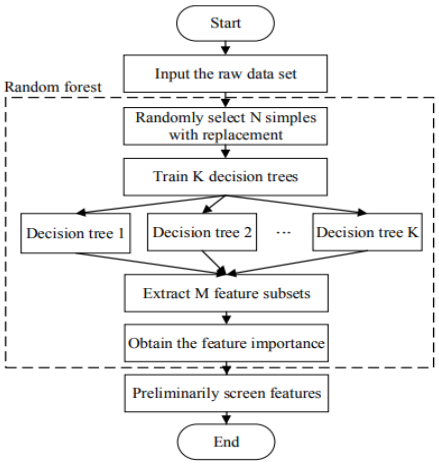


Fig 1: workflow for Random Forest Model

Because each decision tree selection's features are chosen at random from among all M features, the model won't be influenced by a specific feature value or combination of features, increasing randomness and lowering the risk and tendency of overfitting.

By applying specific procedures on the sample classification findings, one can determine the significance of each characteristic and the extent to which it influences the prediction outcome. In the random forest approach, a feature's importance is determined by averaging its importance across all internal decision trees. A feature's impact on the prediction outcomes increases with its significance, and vice versa.

The robustness and effectiveness of the model were ensured by feature selection in the finished framework for predicting lifestyle-related disorders (LRDs). To find the most pertinent features,

methods such as Random Forest importance rankings, Lasso regression, and Recursive Feature Elimination (RFE) were used. By emphasizing variables like body mass index (BMI), alcohol intake, smoking, and physical activity levels, the framework effectively reduced superfluous variables and improved model performance. The selected features enhanced the models' interpretability and increased illness prediction accuracy, allowing medical practitioners to gain a deeper understanding of the underlying risk variables. A simplified dataset that was simpler to handle and produced more precise predictions was the outcome of this successful feature selection procedure.

Due to patient noncompliance, data recording problems, or inadequate surveys, medical databases frequently contain missing or noisy data. In order to create trustworthy models, this problem must be resolved. Missing values can be efficiently filled using imputation techniques like mean, median, mode replacement, or sophisticated algorithms like multiple imputations by chained equations (MICE) and K-Nearest Neighbors (KNN). Additionally, standardization and normalization are used to scale data consistently, especially when features, such daily caloric intake and waist circumference, have different ranges.

In order to ensure that the training data is devoid of abnormalities that could negatively impact model performance, outlier detection and removal techniques like Z-score analysis and interquartile range (IQR) are also essential. By improving the quality of the data, the suggested architecture guarantees reliable input for later.

Datasets:

Parameters including body mass index (BMI), frequency of smoking, alcohol consumption, junk food consumption, regular exercise routines, and waist circumference are frequently included in datasets used to predict lifestyle diseases. A strong basis for predictive modeling is offered by Kaggle, where the labelled dataset used in this study was obtained.

In data mining, feature selection—also known as variable selection—is an essential preprocessing procedure that seeks to improve datasets by eliminating unnecessary and redundant components. In addition to simplifying the data, this procedure increases the effectiveness of machine learning algorithms by enhancing their interpretability, training speed, and predictive accuracy. In healthcare datasets, where noise and duplication are frequent problems, feature selection techniques are especially helpful.

A number of feature selection approaches are frequently used, such as ensemble methods (like Random Forest feature importance), wrapper methods (like recursive feature elimination), filter methods (like correlation-based selection), and embedding methods (like Lasso regression). While eliminating elements that don't significantly improve the model's performance, these techniques methodically find and keep the features that are most pertinent to the prediction objective.

In healthcare applications, reducing the dimensionality of a dataset is essential to handling the high level of complexity of medical data. Predictive models may overfit or converge more slowly if features that are redundant or unnecessary are included in the majority of healthcare datasets. In addition to making modeling easier, good feature selection guarantees that the most clinically important characteristics are given priority, producing quicker, more accurate, and easier-to-understand results.

The majority of current research, according to a study of the literature, focuses on single-disease prediction (91%). Few studies make an effort to forecast several diseases at once. For example:

One-to-One Relationship: The majority of research focuses on utilizing a specific model to forecast a single disease.

One-to-Many Relationship: Only a small number of studies have looked into using the same model and dataset to predict many diseases. There were just three articles that used this strategy.

Many-to-Many Relationship: This method, which is rarely investigated, predicts several diseases using different models across diverse datasets. This approach was only used in one study.

The limitation of each article, as mentioned before, is the defect and the key to forcing the algorithm to perform better. Therefore, we propose and execute a framework for the early diagnosis of depression, diabetes, and hypertension based on lifestyle context, wherein performance assessment criteria were applied to a model efficiency analysis.

The shortcomings of single-disease prediction models highlight the necessity of a more all-encompassing strategy to address diseases linked to lifestyle choices. There is a rising need for frameworks that can forecast many illnesses at once, even though many research concentrate on single conditions like diabetes, hypertension, or obesity. A framework like this would offer a comprehensive viewpoint, facilitating improved management and preventative tactics.

By creating a system that uses feature selection approaches to extract crucial characteristics from redundant and noisy data, our study fills these gaps. The framework seeks to provide accurate and comprehensible predictions for a variety of illnesses, including diabetes, hypertension, and depression, by concentrating on important predictors including daily routines and anthropometric measurements.

This method opens the door for useful applications in healthcare by improving the model's accuracy and robustness while simultaneously reducing its complexity. The focus on feature selection guarantees the framework's continued effectiveness and interpretability, making it a useful instrument for the early identification and treatment of diseases linked to lifestyle choices.

The key features of many diseases are not always precisely identified by current research on predicting Lifestyle-Related Diseases (LRDs). This restriction results from the difficulties of managing dirty, real-world medical datasets and the difficulty of creating reliable models that can handle noisy data. Our objective is to provide a sophisticated and all-encompassing framework for predicting LRD risk. In addition to identifying the most important risk factors from incomplete medical data, our framework will accurately estimate the likelihood of LRDs and visually provide prediction results in a way that is easy to understand.

The framework's three main goals are intended to help address these issues:

Accurate Disease Prediction: Use cutting-edge machine learning and deep learning approaches to create models that produce trustworthy predictions for a variety of LRDs.

Essential Feature Identification: To extract and highlight the most pertinent features from the dataset, use feature engineering and selection techniques.

Analysis and Remediation Strategies: Assist medical professionals in assessing illness risks and suggesting appropriate interventions by offering actionable insights and visualization tools.

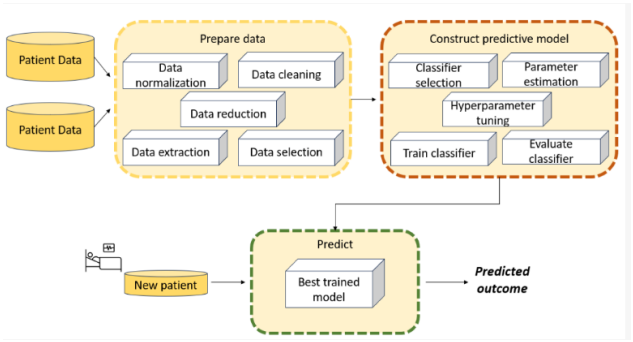


Fig:2 Lifestyle Disease Prediction Model

An outline of the suggested NNSVM algorithm using Random Forest feature selection:

The project's dataset includes everyday lifestyle factors like age, smoking and alcohol consumption, physical activity, stress and anxiety levels, and body mass index (BMI). Because the data is tagged and comes from Kaggle, it can be used for supervised learning tasks. Random Forest Feature Selection (RFFS) is used to extract the most important features from the dataset in order to guarantee the efficacy of prediction models. Predictive modeling relies heavily on feature selection since it streamlines the dataset by eliminating superfluous or irrelevant characteristics, increasing the models' accuracy, efficiency, and interpretability. Diabetes, high blood pressure, and depression are the three noncommunicable disorders that are the subject of this initiative. The intricacy and diversity of these illnesses make it more difficult to choose pertinent qualities. Nevertheless, the RFFS approach makes use of Random Forest's built-in characteristics to prioritize features, making it easier to choose the most pertinent predictors.

Outliers, noise, discrepancies, and missing numbers are all addressed at the data processing stage. Data cleansing, normalization, feature scaling, categorical variable encoding, and dimensionality reduction are examples of data preprocessing methods.

To improve SVM with neural Network called as NNSVM model performance, feature engineering entails choosing, modifying, and producing new features from raw data. This could entail integrating several features, creating new features with domain expertise, or extracting significant features from unprocessed data. In order to create a model that can recognize patterns and connections between input attributes and output labels, data modeling is an essential part of the architecture. To create precise and reliable machine learning systems, effective data modeling necessitates careful consideration of feature engineering, model selection, data preprocessing, and evaluation methodologies. The last step is model evaluation. Once a model has been trained, its performance on unseen data is assessed using a validation set. After that, the best model is selected from the evaluated models.

It can be costly and difficult to obtain properly labeled datasets, especially in domains like medical imaging where lengthy annotation procedures or large studies may be necessary. Consequently, the capacity to efficiently learn from data with poor labels has grown in significance. Based on the type of labels, weak-labeling problems are divided into three categories: implicitly known labels (where

samples are organized into labeled sets or "bags"), partially known labels (where the majority of training samples lack labels), and fully unknown labels.

Label and feature noise, among other problems with data quality, make these difficulties even more difficult. These problems can have serious repercussions in medical applications, since diagnostic tests are not always entirely correct. Specifically, label noise can impair classifier performance, raising the learning difficulty and resource needs. Furthermore, it has the potential to skew the observed frequency of medical test results, which could result in incorrect inferences about a population's characteristics. These difficulties highlight how crucial it is to create reliable techniques for efficiently handling noisy and poorly labeled data.

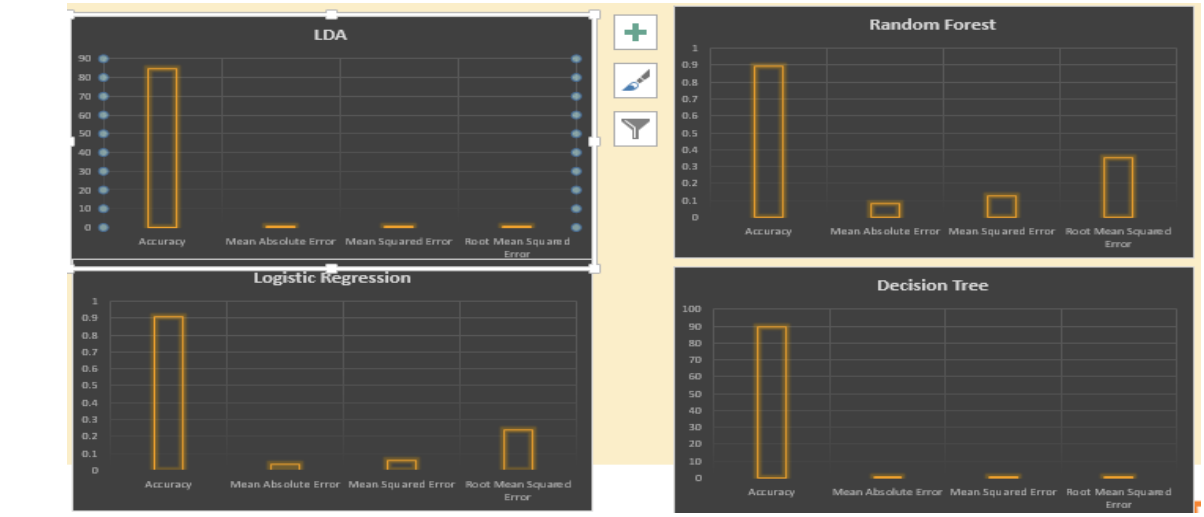
Performance Analysis:

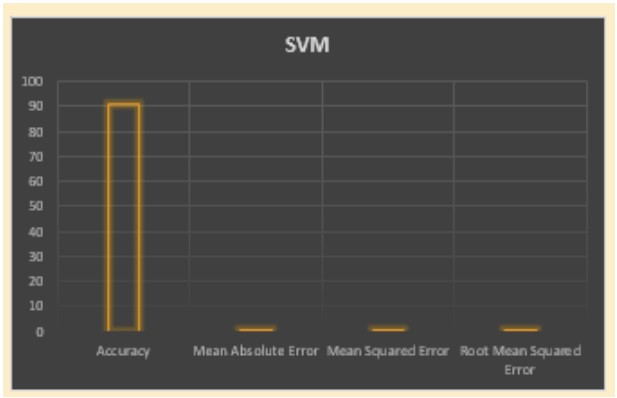
Performance measures are crucial for assessing the dependability and efficacy of machine learning models, especially when it comes to the prediction of lifestyle diseases. These metrics offer a numerical assessment of a model's ability to forecast results from unknown data. Precision, which shows the percentage of correctly predicted positive cases out of all predicted positives, and accuracy, which calculates the overall percentage of correctly classified instances, are often used metrics. The model's capacity to detect all real positive cases is measured by recall, often referred to as sensitivity. The F1-score, which combines precision and recall into a single metric, offers a balanced measure that is especially helpful for datasets that are unbalanced. Furthermore, the model's discriminative strength is highlighted by the ROC-AUC (Receiver Operating Characteristic - Area Under Curve) metric, which assesses the model's capacity to differentiate between classes across a range of thresholds. Additional metrics, such as specificity, evaluate how well the model detects negative cases. Metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared can be used to gauge how accurate predictions are for regression tasks. Together, these performance measures aid in assessing the model's resilience, dependability, and usefulness in real-world situations, directing future optimization and deployment choices.

3. RESULT AND DISCUSSION:

The Random Forest Feature Selection Model and the NNSVM (Neural Network Support Vector Machine) approach are implemented in this study. Creating a feature selection model and examining how it affects disease prediction are the main goals. Three lifestyle-related diseases—depression, diabetes, and hypertension—are specifically highlighted in the study. In order to determine baseline accuracies, a number of machine learning models were first developed without the use of any feature selection algorithms. These models included Linear Discriminant Analysis, Random Forest, Logistic Regression, Decision Tree, and Support Vector Machine. The purpose of this stage was to assess how well these models performed when all of the characteristics were utilized without any refining. Each model's accuracy was noted and examined to demonstrate how important feature selection is for enhancing prediction results. The study illustrates how choosing the most pertinent features improves the effectiveness and precision of disease prediction models by contrasting outcomes before and after using the Random Forest Feature Selection Model. Here all the features are considered and results are compared with the different ML Algorithms.The results of the above algorithms are as in Fig. 3.

Now after applying Random Forest feature selection algorithm some important features are extracted. These features are further used to implement same ML algorithms and these features have improved the accuracy of the models.





Sr. No.	ML Algorithms	Accuracy
1	LDA	84.94983278
2	SVM	91.30435
3	Decision Tree	89.63210702
4	Random Forest	89.6454849
5	Logistic Regression	91.2244147

Fig. 3 Machine Learning Algorithms and Accuracy Comparison

A subset of the original dataset's most significant and pertinent features were found when the Random Forest feature selection algorithm was used. Given their strong associations with the desired results, these particular characteristics were thought to be essential for forecasting lifestyle disorders including depression, diabetes, and hypertension. The same machine learning methods, including Linear Discriminant Analysis, Random Forest, Logistic Regression, Decision Tree, and Support Vector Machine, were retrained using this improved dataset following the extraction of these crucial characteristics.

The models were able to produce more accurate forecasts by concentrating just on the most pertinent features. By removing unnecessary or redundant variables that would have hampered the model's performance, the feature selection procedure successfully decreased noise. Because the algorithms were able to concentrate on the crucial data patterns that have a direct impact on the outcomes of the disease, the models trained using the chosen features shown a noticeable increase in accuracy.

This improvement emphasizes how important feature selection is in machine learning processes, particularly when dealing with healthcare datasets, where noisy or irrelevant data might impair the

model's ability to predict outcomes. Only the most significant features were chosen, which improved the models' speed, efficiency, and ability to produce more accurate predictions for lifestyle-related

Analysis of Feature Selection

A key factor in improving the overall effectiveness of prediction models is feature selection. To increase the precision and effectiveness of the prediction models, a hybrid NNSVM (Neural Network Support Vector Machine) technique is used in this suggested work. Cross-validation is used to assess many feature subsets and choose the one with the highest predictive value in order to determine which attributes are most pertinent. This procedure lowers noise and increases prediction accuracy by ensuring that only the most important features are included for model training.

1. BMI
2. Age
3. Smoking
4. Alcoholism
5. Stress and Anxiety
6. lack of Physical Activity

Predicting the three lifestyle-related disorders of depression, diabetes, and hypertension is the main goal of this study. The model can learn from the most significant data points by utilizing the NNSVM approach in conjunction with feature selection, producing predictions that are more robust and dependable. Hybrid NNSVM and feature selection work together to improve the model's performance and increase its ability to predict these diseases' risk based on pertinent lifestyle factors.

The best features chosen by the Random Forest Feature Selection (RF FE) model are used to apply machine learning algorithms once more, and the outcomes are compared. The contrast between the machine learning models' performance with and without feature selection is depicted in the figure. This comparison shows the substantial influence of feature selection on predictive performance by highlighting how the addition of specific features enhances the models' efficiency and accuracy.

Sr. No.	ML Algorithms	Accuracy	Accuracy After Feature Selection
1	LDA	84.94983278	89.634524
2	SVM	91.30435	93.567404
3	Decision Tree	89.63210702	92.543832
4	Random Forest	89.6454849	91.07689
5	Logistic Regression	91.3244147	93.6744043

Fig.4. Comparison of the results with and without Feature selection

Initially, a range of classification strategies are investigated and used to enhance the effectiveness of machine learning algorithms in disease prediction. These include widely used techniques like Support Vector Machines (SVM), Random Forest, and Decision Trees. Following Random Forest feature selection, these classifiers' efficacy is evaluated by comparing their precision, prediction accuracy, and other pertinent performance indicators. Finding the best categorization method for the given problem is made possible by this comparison study. According to the evaluation's results, Support Vector Machines (SVM) routinely perform better than the alternative techniques in terms of classification accuracy.

In light of SVM's encouraging performance, a hybrid strategy is suggested to improve predicted accuracy even more accurate and dependable predictions because to this ensemble approach, which capitalizes on the various advantages of each technique.

Sr. No.	Algorithms	Accuracy obtained with the different algorithms
1	LDA	89.634524
2	SVM	93.567404
3	Decision Tree	92.543832
4	Random Forest	91.07689
5	Logistic Regression	93.6744043
6	**NNSVM Ensemble(Proposed Methodology)**	96.65543

Fig.5 Comparison of NNSVM proposed Algorithm with other Methods

As seen by the findings in Fig 5, where the combined model performs better than the individual models, ANNs and SVMs can be integrated to provide a notable improvement in prediction performance. The model is better able to manage the intricacies of forecasting lifestyle-related illnesses by employing this hybrid approach, which eventually raises the precision of early detection and diagnosis in medical applications. After a careful examination of the data gathered from the different methods, the NNSVM Method performs better than the others. The diseases caused by specific lifestyle choices can then be treated with our recommendations.

4. CONCLUSUION:

This study examines several disease prediction algorithms and assesses how accurate they are under various conditions. It is evident from a survey of the literature that the majority of prediction models base their diagnosis mostly on patient symptoms. Predictive models that take lifestyle factors like nutrition, exercise, smoking, and alcohol use into account, however, are conspicuously lacking. Chronic diseases including depression, diabetes, and hypertension are significantly influenced by these characteristics, however it can be difficult to locate datasets that include these lifestyle factors. We fill this gap by implementing a unique hybrid model termed NNSVM, which improves predictive accuracy by combining Artificial Neural Networks (ANNs) with Support Vector Machines (SVMs).

Because ANNs can capture complicated, non-linear correlations in the data, they are excellent at learning complex patterns from huge, high-dimensional datasets. Because of this, they are especially good at identifying patterns that conventional models might not be able to show right away. SVMs,

on the other hand, are well-known for their capacity to locate the best decision boundaries in high-dimensional data, which makes them the perfect option for classification problems. The hybrid NNSVM model combines the qualities of both approaches by fusing the potent classification capacity of SVMs with the feature learning capabilities of ANNs. It is anticipated that this integrated approach will produce a more dependable model, improving early detection of lifestyle-related health issues and increasing predicting accuracy for diseases. The variety of lifestyle characteristics that are incorporated into the prediction model may be increased in further research. Other variables, like sleep patterns, stress levels, social determinants of health, and mental health, could be included in addition to BMI, smoking, drinking, and physical activity. The model's capacity to forecast a greater range of illnesses might be enhanced by taking into account a more extensive collection of lifestyle factors.

REFERENCE

[1] Trends in coronary Heart Disease Epidemiology
[2] Center for Disease Control and Prevention (Heart Disease Facts).
[3] Asian Pacific Journal of Global Trend of Cancer Mortality rate: A 25-year study.
[4] Times Of India: Cancer cases upswing 10% in 4 years to 13.9 lakh.
[5] International Diabetes Federation: Expenditure and deaths related to diabetes.
[6] Epidemiology of Diabetes :A report of Indian Heart Association.
[7] Naveen Kishore G,V .Rajesh ,A.Vamsi Akki Reddy, K.Sumedh,T.rajesh Sai Reddy, ”Prediction Of Diabetes Using Machine Learning Classification Algorithms”.
[8] Gavin Pinto, Sunil Jangid, Radhika Desai, ”Understanding the Lifestyle of people to identify the reasons of Diabetes using data mining”.
[9] M.Marimuthu ,S.Deivarani ,R.Gayatri, “Analysis of Heart Disease Prediction using Machine Learning Techniques”.
[10] Purushottam, Richa Sharma ,Dr. Kanak Saxena, ”Efficient Heart Disease Prediction System”.
[11] Adil Hussain She, Dr. Pawan Kumar Chaurasia,” A Review on Heart Disease Prediction using Machine Learning Techniques”.
[12] M. Chinna Rao ,K. Ramesh, G. Subbalakshmi,”Decision Support in Heart Disease Prediction System using Naïve Bayes”.
[13] Amandeep Kaur , Jyothi Arora,” Heart Disease Prediction using data mining Techniques :A survey”.
[14] Noreen Fatima , Li Liu , Sha Hong, Haroon Ahmed ,”Prediction of Breast Cancer, Comparitive Review Of Machine Learning Algorithms and their analysis”.
[15] Ch .Shravya ,K.Pravallika , Shaik Subhani, ”Prediction of Cancer using supervised machine learning Algorithms”.
[16] Nikita Rane, Jean Sunny, Rucha Kanade, Sulochana Devi,” Breast Cancer classification and prediction using machine learning “.
[17] Deepti Sisodia, Dilip Singh Sisodia,” Prediction of Diabetes using classification Techniques”.
[18] Dr.B.Santhosh Kumar, T.Daniya, Dr. J.Ajayan,” Breast Cancer Prediction using Machine Learning Algorithms”. [19] Mümine KAYA KELEŞ ,”Cancer Prediction using and Detection using Machine Learning Algorithms : A Comparitive Study”.
[20] Heart Disease Dataset” by UCI.
[21] B.-H. Chew, H. Hussain, and Z. A. Supian, ‘‘Is therapeutic inertia present in hyperglycaemia, hypertension and hypercholesterolaemia management among adults with type 2 diabetes in three health clinics in malaysia? A retrospective cohort study,’’ BMC Family Pract., vol. 22, no. 1, p. 111, Dec. 2021, doi: 10.1186/s12875-021-01472-2. [22] H. S. A. Fang, Q. Gao, W. Y. Tan, M. L. Lee, W. Hsu, and N. C. Tan, ‘‘The effect of oral diabetes medications on glycated haemoglobin (HbA1c) in Asians in primary care: A retrospective cohort real-world data study,’’ BMC Med., vol. 20, no. 1, p. 22, Dec. 2022, doi: 10.1186/s12916-021-02221-z.
[23] K. Donsa, S. Spat, P. Beck, T. R. Pieber, and A. Holzinger, ‘‘Towards personalization of diabetes therapy using computerized decision support and machine learning: Some open problems and challenges,’’ in Smart Health: Open

Problems and Future Challenges, A. Holzinger, C. Röcker, and M. Ziefle, Eds. Cham, Switzerland: Springer, 2015, pp. 237–260.

[24] S. Ellahham, ‘‘Artificial intelligence: The future for diabetes care,’’ Amer. J. Med., vol. 133, no. 8, pp. 895–900, 2020, doi: 10.1016/j.amjmed.2020.03.033.

[25] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, ‘‘Machine learning and data mining methods in diabetes research,’’ Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, Jan. 2017, doi: 10.1016/j.csbj.2016.12.005.

[26] Guido, R.; Ferrisi, S.; Lofaro, D.; Conforti, D. An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. *Information* 2024, *15*, 235. https://doi.org/10.3390/info15040235

[27] Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, New York, NY, USA, 27–29 July 1992; COLT’92. pp. 144–152. [Google Scholar] [CrossRef]

[28] Cortes, C.; Vapnik, V. Support-vector networks. Mach. Learn. 1995, 20, 273–297. [Google Scholar] [CrossRef] Vapnik, V. An overview of statistical learning theory. IEEE Trans. Neural Netw. 1999, 10, 988–999. [Google Scholar] [CrossRef] [PubMed]

[29] Burges, C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Min. Knowl. Discov. 1998, 2, 121–167. [Google Scholar] [CrossRef]