Journal of Advanced Zoology
ISSN: 0253-7214
Volume 45 Issue S-4 Year 2024 Page 65-72

A Comprehensive Review On The Analysis Of Various Machine Learning Algorithms For Early Detection Of Critical Diseases

Divya Chitre[1*], Prof. Shivendu Bhushan[2], Dr. Manisha S Patil[3]

[1,* 2,3] *Indira college of commerce and Science.*
Email: divya.chitre@iccs.ac.in[1], shivendu@iccs.ac.in[2], manisha.patil@issc.ac.in[3]

[*]***Correspondence Author:*** *Divya Chitre*
[*] *Indira college of commerce and Science.*
Email: divya.chitre@iccs.ac.in

Abstract

Early detection of critical diseases is a pivotal aspect of modern healthcare, significantly impacting patient outcomes and healthcare costs. This research paper provides a comprehensive review and analysis of various machine learning algorithms employed in the realm of early disease detection. The study explores the strengths, limitations, and overall efficacy of prominent algorithms, including Logistic Regression, Support Vector Machines, Random Forests, Neural Networks, K-Nearest Neighbors, and Ensemble Learning. Each algorithm's suitability for early detection is assessed based on factors such as interpretability, scalability, and performance in handling diverse data types. Furthermore, the review discusses the specific applications of these algorithms in different medical contexts, highlighting their contributions to the early identification of critical diseases. By synthesizing the current state of research, this paper aims to provide valuable insights for researchers, and policymakers working towards advancing the field of early disease detection through machine learning.

Keywords: Early Detection, Critical Diseases, Machine Learning Algorithms.

INTRODUCTION:

Within the dynamic realm of healthcare, the timely identification of pivotal diseases like Cancer, Cardiovascular Diseases, Diabetes, Alzheimer's, and Chronic Kidney Disease (CKD) serves as a foundational pillar, aiming to enhance patient outcomes and alleviate the strain on healthcare systems. This imperative focus on early detection underscores the significance of proactive measures in addressing critical health conditions, paving the way for more effective treatments and ultimately contributing to a more sustainable and responsive healthcare landscape.

Timely identification of diseases at their nascent stages not only facilitates more effective treatment but also holds the potential to enhance the overall quality of life for affected individuals. In this context, the integration of machine learning (ML) algorithms has emerged as a promising avenue, leveraging the power of computational intelligence to sift through vast and complex datasets for subtle patterns indicative of impending

health challenges. This research paper undertakes a comprehensive review and analysis of diverse ML algorithms employed in the critical domain of early disease detection. By scrutinizing the strengths, limitations, and performance metrics of these algorithms, this study aims to provide a nuanced understanding of their applicability and efficacy in the pursuit of timely and accurate identification of critical diseases. As we embark on this exploration, we seek not only to elucidate the current landscape but also to pave the way for future advancements in leveraging machine learning for the early detection of diseases that demand swift intervention.

LITERATURE REVIEW:

There have been numerous studies done related to predicting the disease using different machine learning techniques and algorithms which can be used by medical institutions. This paper reviews some of those studies done in research papers using the techniques and results used by them.
Chetty et al [1] proposed a system utilizing a fuzzy approach for disease prediction. The study incorporated machine learning techniques, including the KNN classifier, Fuzzy c-means clustering, and Fuzzy KNN classifier. The accuracy achieved for predicting diabetes was 97.02%, and for liver disorders, it was 96.13%.
Shukla et al [2] explored the prediction and detection of breast cancer using various machine learning techniques such as Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network, and KNN. Their findings highlighted the superior accuracy of the Support Vector Machine in comparison to other algorithms.

Chen et al [3] proposed a disease prediction system employing machine learning algorithms, including CNN-UDRP and CNN-MDRP algorithms, Naive Bayes, K-Nearest Neighbor, and Decision Tree. The accuracy achieved by their system was reported as 94.8%
Ambekar et al [4] recommended disease risk prediction using a convolutional neural network, alongside machine learning techniques such as CNN-UDRP, Naive Bayes, and KNN. The structured data-driven approach resulted in an accuracy of 82%, with Naïve Bayes contributing significantly.
Lambodar Jena et al Paper Title: "Risk Prediction for Chronic Diseases using Distributed Machine Learning Classifiers"
Jena et al [5] focused on risk prediction for chronic diseases, specifically Chronic Kidney Disease. They employed distributed machine learning classifiers, achieving high accuracies of 95% and 99.7% with Naïve Bayes and Multilayer Perceptron, respectively [Reference:

Pahulpreet Singh Kohli et al, [6] suggested disease prediction by using applications and methods of machine learning and used techniques like Logistic Regression, Decision Tree, Support Vector Machine, Random Forest and Adaptive Boosting. This paper focuses on predicting Heart disease, Breast cancer, and Diabetes. The highest accuracies are obtained using Logistic Regression that is 95.71% for Breast cancer, 84.42% for Diabetes, and 87.12% for Heart disease.
Deeraj Shetty et al, [7] studied the uses of data mining for diabetes disease prediction by using Naïve Bayes and KNN algorithms. This system predicts diabetes and accuracy obtained by KNN are better than Naïve Bayes.
Several research studies have explored the application of various machine learning techniques for disease prediction, offering valuable insights for medical institutions.

SUPERVISED LEARNING FOR EARLY DETECTION:

Labeled Data: Supervised learning requires labeled data, where instances are paired with corresponding labels indicating the presence or absence of the disease. In the context of early disease detection, having historical data with known outcomes (positive or negative cases) is essential.
Training for Specific Patterns: Supervised algorithms learn patterns and relationships in the labeled data, enabling them to make predictions on new, unseen data. This is crucial for early detection, as the algorithm can recognize patterns associated with the early stages of a disease based on the training data.

Common supervised learning algorithms for early disease detection include Logistic Regression, Support Vector Machines (SVM), Random Forest, and Neural Networks. These algorithms can be trained on labeled datasets containing features relevant to early disease indicators.

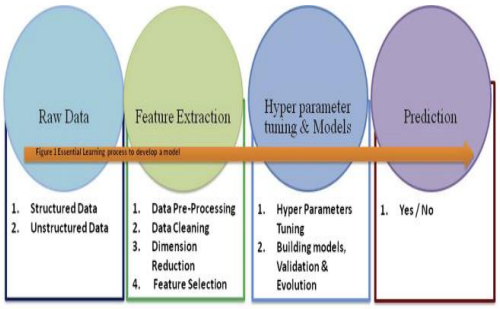

Figure 1 Essential Learning process to develop predictive model [10]

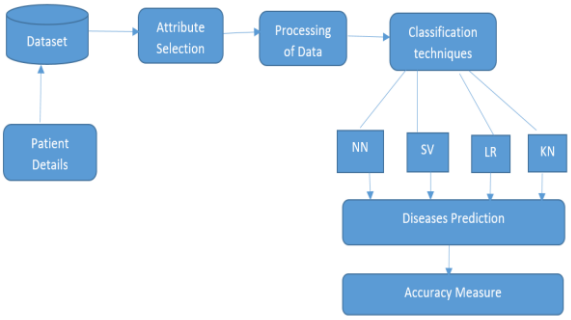

Figure2 The Architecture Of Disease Prediction System

SUPPORT VECTOR MACHINES (SVM):

The Support Vector Machine (SVM) serves as a versatile tool applicable to both classification and regression tasks. In the SVM model, data points are represented in a multidimensional space and are grouped based on similarities, where points with akin characteristics belong to the same category.

In the context of linear SVM, the dataset is conceptualized as a p-dimensional vector space, and the objective is to identify a hyperplane that maximally separates the data into distinct groups. These hyperplanes, numbering up to p-1, act as boundaries in the data space, defining the limits between different categories in classification or regression problems.

The selection of the optimal hyperplane involves considering the distances between classes. Among the various hyperplanes, the one with the maximum margin, referred to as the maximum-margin hyperplane, is chosen. This margin signifies the perpendicular distance between the hyperplane and the nearest data points of the two classes. In essence, linear SVM seeks to establish the most effective hyperplane for class separation by maximizing the margin between different classes. This approach enhances the model's robustness and generalization ability, making it capable of accurately classifying or predicting new data points based on their position relative to the identified hyperplane. Here are some of the key benefits of using Support vector Machine.

Classification Prowess:

SVM emerges as a preeminent classification algorithm, adeptly handling binary or multiclass tasks. In the domain of disease detection, SVM proves instrumental in training models to categorize individuals based on pertinent features, distinguishing between health and disease states.

Navigating High-Dimensional Spaces:

Diseases often manifest through intricate interactions among multiple variables. SVM's effectiveness in high-dimensional spaces positions it as a fitting choice for scenarios where a myriad of features contribute to the complexities of disease detection.

Flexibility in Capturing Non-Linear Relationships:

SVM's utilization of diverse kernel functions facilitates the mapping of input data into higher-dimensional spaces, enabling the capture of non-linear relationships among features. This flexibility proves invaluable when unraveling intricate patterns and associations embedded in medical data.

Quest for the Optimal Hyperplane:
At its core, SVM operates by identifying the optimal hyperplane that maximally separates data points of different classes. This emphasis on margin maximization empowers SVM to generalize effectively to new, unseen data—a critical attribute in the realm of early disease detection.

Addressing Imbalanced Datasets:
Medical datasets often grapple with imbalances between healthy and diseased cases. SVM's capability to focus on correctly classifying instances on the margins renders it well-equipped to handle imbalanced datasets. This becomes particularly pertinent in the context of early detection, where positive cases may be scarce.

Unveiling Feature Significance:
A distinctive strength of SVM lies in its provision of insights into feature importance through support vectors. This feature proves invaluable for clinicians, aiding in the interpretation of model findings and identification of potential early indicators of diseases.

Versatility Across Medical Domains:
SVM's prowess extends across various medical domains, finding successful applications in cancer diagnosis, diabetes prediction, and the detection of neurological disorders. Its adaptability renders SVM a versatile tool applicable to a wide spectrum of diseases and conditions.

RANDOM FOREST (RF):
The random forest (RF) algorithm is a sophisticated ensemble of tree-structured base classifiers. Typically applied to text data with numerous dimensions, such datasets often contain a surplus of irrelevant attributes, with only a handful being crucial for an effective classifier model. The RF algorithm employs a predetermined probability to select the most significant and relevant attributes.

The formulation of the RF algorithm by Breiman involves sampling data subsets and constructing multiple decision trees by mapping random samples of feature subspaces. The RF algorithm, associated with a set of training documents (D) and Nf features, can be outlined as follows:

Initialization: Samples D1, D2, …, DK are chosen with replacement based on predetermined probabilities.

Tree Construction: For each document DK, a decision tree model is constructed. The training documents are randomly sampled using a subspace of m-try dimensions from the available features. Probabilities are calculated based on the m-try features, and the leaf node determines the optimal data split. This process continues until a saturation criterion is met.

The K unpruned trees (h1(X1), h2(X2), …) are then combined into a random forest ensemble, and the high probability value is utilized for classification decisions.

Random Forest Pseudocode:
- Randomly select "n" features from the total "k" features, where n << k.
- Calculate the node "n" among the selected features using the best split point.
- Categorize the node into daughter nodes using the optimal split.
- Repeat steps 1 to 3 until "l" nodes are reached.
- Build the forest by repeating steps 1 to 4 "n" times to create "n" trees.

In simpler terms, the random forest algorithm involves selecting a subset of features, determining the best way to split nodes based on these features, and iteratively constructing decision trees. The final random forest ensemble is created by repeating this process multiple B times, resulting in a robust and accurate classification model. Benefits of employing Random Forest in this critical domain.

Ensemble Learning Excellence:
Random Forest excels in the concept of ensemble learning, where it constructs multiple decision trees and combines their outputs. This ensemble approach enhances the model's predictive performance, making it well-suited for nuanced tasks such as early disease detection.

Versatility in Classification:
The algorithm's primary forte lies in classification tasks, making it highly applicable for categorizing individuals into distinct groups, such as healthy or diseased. Its ability to handle both binary and multiclass classification scenarios contributes to its versatility in medical applications.

Accommodating High-Dimensional Data:
Diseases often involve complex interactions among numerous factors and variables. Random Forest's inherent capability to handle high-dimensional data is crucial in scenarios where a multitude of features contribute to the early detection of diseases.

Resilience to Overfitting:
Random Forest mitigates the risk of overfitting, a common concern in machine learning. By aggregating predictions from multiple trees and incorporating randomness during training, the algorithm maintains a balance between model complexity and generalization, fostering reliable predictions.

Non-Linear Pattern Recognition:
The algorithm's ability to capture non-linear relationships within data is pivotal for early disease detection. Random Forest achieves this by considering subsets of features for each decision tree, collectively providing a more comprehensive understanding of intricate patterns inherent in medical datasets.

Handling Imbalanced Datasets:
In the medical field, datasets often exhibit imbalances between healthy and diseased instances. Random Forest handles such imbalances adeptly, as each decision tree focuses on different subsets of data, ensuring a nuanced approach to classification even in scenarios where positive cases are scarce.

Feature Importance Insights:
Random Forest provides valuable insights into feature importance, allowing healthcare professionals to discern which factors contribute significantly to the early detection of diseases. This transparency aids in understanding the underlying mechanisms and facilitates more informed decision-making.

Applicability Across Medical Domains:
Random Forest's versatility extends across diverse medical domains. It has demonstrated success in various applications, including cancer prediction, cardiovascular disease detection, and identification of neurological disorders. This broad applicability makes Random Forest a valuable asset for healthcare practitioners.

LOGISTIC REGRESSION:
Logistic Regression, despite its name, is a powerful classification algorithm that has proven to be effective in various fields, including healthcare.
Logistic Regression, a technique borrowed from statistics, is a go-to method for binary classification problems. In this context, the logistic function plays a central role. This sigmoid function, initially developed by statisticians to describe population growth in ecology, transforms real-valued numbers into a range between 0 and 1, resembling an S-shaped curve.
The logistic function is expressed as 1/(1+e−value), where 'e' is Euler's number, and 'value' is the numerical input to be transformed. This function ensures values are mapped into the specified range, but never exactly at the limits of 0 or 1.
In logistic regression, the logistic function is utilized to model the relationship between input features and the binary output (0 or 1). The representation involves an equation similar to linear regression, where input values (x) are linearly combined using weights or coefficients (denoted as Beta) to predict the output value (y). Notably, the output value in logistic regression is binary, distinguishing it from linear regression.
The logistic regression equation can be illustrated as:

$$y = \frac{e^{(b0+b1 \cdot x)}}{1+e^{(b0+b1 \cdot x)}}$$

Here, y represents the predicted output, b_0 is the bias or intercept term, and b_1 is the coefficient associated with the input value x. Each column in the input data corresponds to a distinct b coefficient, and these coefficients are learned from the training data. Benefits of employing Logistic Regression in medical domain are:-

Probabilistic Classification:
Logistic Regression is particularly well-suited for binary classification tasks, making it applicable to scenarios where the objective is to categorize individuals into two groups, such as healthy or diseased. The algorithm estimates the probability of an individual belonging to a particular class, providing a nuanced approach to disease detection.

Interpretability and Transparency:
One of the strengths of Logistic Regression lies in its interpretability. The model outputs coefficients for each feature, allowing healthcare professionals to understand the influence of different variables on the likelihood of disease. This transparency is crucial for gaining insights into the early indicators identified by the model.

Handling High-Dimensional Data:
In healthcare, datasets often involve numerous features that may contribute to the early detection of diseases. Logistic Regression can handle high-dimensional data, offering a straightforward approach to modeling complex relationships between features and the likelihood of disease occurrence.

Resilience to Overfitting:
Logistic Regression is less prone to overfitting compared to more complex models, which can be advantageous in scenarios with limited data for training. Its simplicity ensures a balance between model complexity and generalization, making it robust for early disease detection tasks.

Probability Threshold Adjustment:
Logistic Regression outputs probabilities, and practitioners can adjust the probability threshold to customize the trade-off between sensitivity and specificity based on the specific needs of the early detection task. This flexibility allows for fine-tuning the model's performance to align with clinical requirements.

Scalability and Computational Efficiency:
Logistic Regression is computationally efficient and scalable, making it suitable for handling large healthcare datasets. Its efficiency is particularly valuable when working with real-time or high-throughput data streams, essential in the context of early disease detection.

Application in Risk Stratification:
Logistic Regression is commonly used in risk stratification, where individuals are categorized into different risk groups based on their likelihood of developing a particular disease. This risk assessment is foundational for early detection efforts, guiding preventive interventions for high-risk individuals.

Real-world Applications:
Logistic Regression has been successfully applied in various medical domains, including cardiovascular disease risk prediction, diabetes diagnosis, and cancer detection. Its versatility makes it an accessible and effective tool for healthcare practitioners addressing diverse early detection challenges.

K-NEAREST NEIGHBORS (KNN):
The k-Nearest Neighbors (k-NN) algorithm is a straightforward yet effective method that falls under the category of lazy, non-parametric, and instance-based learning. It is versatile and applicable to both classification and regression tasks. In classification scenarios, k-NN is employed to determine the class to which a new, unlabeled object belongs. The process involves selecting a value for 'k,' typically an odd number, which represents the number of neighbors to consider. Distances between the data points, often calculated using metrics like Euclidean distance, Hamming distance, Manhattan distance, or Minkowski distance, are computed to identify the nearest neighbors to the new object. Once the distances are determined, the 'k' nearest neighbors are selected, and the class of the new object is predicted based on the majority votes or consensus of these neighbors. In other words, each neighbor "votes" for its respective class, and the class with the highest number of votes becomes the predicted class for the new object. Overall, the k-NN algorithm provides accurate

predictions by leveraging the collective information from the neighboring data points, making it a powerful and simple tool for classification tasks here, benefits of using KNN in the context of early disease detection:

Proximity-Based Classification:
KNN operates on the principle of proximity, classifying instances based on the majority class of their k-nearest neighbors. In the context of early disease detection, this approach can be valuable when the patterns indicating disease presence are reflected in the proximity of feature values.

Simplicity and Intuitiveness:
KNN is inherently simple and intuitive, making it accessible for healthcare professionals and practitioners. Its straightforward approach is beneficial when dealing with early detection tasks, as it allows for easy interpretation of results and decision-making.

No Assumptions about Data Distribution:
KNN makes minimal assumptions about the underlying data distribution, making it suitable for scenarios where the nature of the relationship between features and disease indicators may not be explicitly known. This flexibility is advantageous in early detection efforts where comprehensive understanding is evolving.

Adaptability to Feature Changes:
As new features or variables become relevant in the context of early detection, KNN can easily adapt to these changes. The algorithm doesn't require retraining when additional features are introduced, providing a dynamic and flexible framework for evolving healthcare data.

Sensitivity to Local Patterns:
KNN is sensitive to local patterns in the data, allowing it to capture intricacies that may vary across different regions of the feature space. This sensitivity can be advantageous in early disease detection, where localized patterns might be indicative of specific stages or manifestations of a disease.

Consideration of Feature Importance:
While KNN doesn't explicitly provide feature importance like some other algorithms, the proximity-based classification inherently considers the relevance of each feature based on its contribution to the distances between instances. This aspect can aid in understanding the significance of features in the context of early detection.

Handling Non-Linear Relationships:
KNN is capable of capturing non-linear relationships in the data, as it relies on distances rather than linear decision boundaries. This characteristic is crucial when dealing with complex patterns and interactions between variables in early disease detection scenarios.

Limitations on Large Datasets:
One consideration with KNN is its computational cost, especially as the dataset size increases. For very large healthcare datasets, the algorithm's efficiency might be a factor to consider, and optimizations or alternative algorithms may be explored.

Real-world Applications:
KNN has found applications in healthcare, including disease prediction, risk assessment, and outcome prediction. Its adaptability and simplicity make it suitable for a range of medical scenarios, contributing to the early detection of diseases.

CONCLUSION:

When selecting an algorithm for disease detection, it's also important to consider factors like interpretability, computational efficiency, and the ability to handle imbalanced datasets (common in healthcare). Logistic Regression and SVMs are generally more interpretable compared to ensemble methods like Random Forest, which may be an important consideration in medical applications where interpretability is crucial. In practice, it's advisable to experiment with multiple algorithms, fine-tune their parameters, and evaluate their

performance using relevant metrics (e.g., sensitivity, specificity, ROC-AUC) on a validation set or through cross-validation. It might also be beneficial to consult with domain experts to ensure that the selected model aligns with the medical knowledge and requirements of the specific disease being targeted.

REFERENCES:

1. Chetty, N., Vaisla, K. S., & Patil, N. (2015, May). An improved method for disease prediction using fuzzy approach. In 2015 Second International Conference on Advances in Computing and Communication Engineering (pp. 568-572). IEEE.
2. Rati Shukla, Vikash Yadav, Parashu Ram Pal and Pankaj Pathak, "Machine Learning Techniques for Detecting and Predicting Breast Cancer" IJITEE, ISSN: 2278-3075, Volume-8, pp. 2658-2662, 2019.
3. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities" IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.
4. Sayali Ambekar, Rashmi Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network" IEEE, 978-1-5386-5257-2/18, 2018
5. Lambodar Jena and Ramakrushna Swain, "Chronic Disease Risk Prediction using Distributed Machine Learning Classifiers" IEEE, 978-1-5386-2924-6/17, pp. 170-173, 2017.
6. Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Disease Prediction" IEEE, 978-1-5386-6947-1/18, pp. 1-4, 2018.
7. Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil, " Diabetes Disease Prediction Using Data Mining" IEEE, 978-1-5090-3294-5/17, 2017.
8. Farooqui, M. E., & Ahmad, D. J. (2020). Disease prediction system using support vector machine and multilinear regression. International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, 2347-5552.
9. Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing, 77, 5198-5219.
10. Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. Applied computing and informatics, 18(1/2), 90-100.
11. Bansal, D., Chhikara, R., Khanna, K., & Gupta, P. (2018). Comparative analysis of various machine learning algorithms for detecting dementia. Procedia computer science, 132, 1497-1502.
12. Alam, M. Z., Rahman, M. S., & Rahman, M. S. (2019). A Random Forest based predictor for medical data classification using feature ranking. Informatics in Medicine Unlocked, 15, 100180.
13. https://www.shethepeople.tv/health/cardiovascular-diseases-among-indian-women-high-the-lancet-report/
14. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6994761