# DIAGNOSIS AND DETECTION OF VARIOUS DISEASES USING MACHINE LEARNING HEURISTICS: A STUDY

**Abstract**

Think about the circumstances of those residing in an area remote from a hospital. Consider the people who lack the resources to cover the hospital bills as well. There are some incredibly busy people in our society who do not have enough time in their lives to visit the hospital for the diagnosis of their own diseases or the diseases of other family members. In the aforementioned scenarios, diseases can be identified using cutting-edge medical technology, perhaps saving our precious lives. Several artificially intelligent diagnosis heuristics have been developed by researchers to differentiate between a number of common ailments. The focus of the chapter is on recent advances in machine learning (ML) that have had a significant impact on the diagnosis and recognition of various illnesses.

**Keywords:** Disease and ML, Diagnosis, Detection, machine learning, Heuristics

**Authors**

**Ira Nath**
Ph. D
Department of Computer Science & Engineering
JIS College of Engineering
Kalyani, West Bengal, India.
ira.nath@gmail.com

**Pranati Rakshit**
Ph. D
Department of Computer Science & Engineering
JIS College of Engineering
Kalyani, India.
pranatirakshit17@gmail.com

**Dharmpal Singh**
Ph. D
Department o of Computer Science & Engineering
JIS University
Agarpara, Kolkata, West Bengal, India.
dharmpal1982@gmail.com

## I. INTRODUCTION

**Introduction of Machine learning:** The process of teaching and learning a machine so that it can generate the correct information is known as machine learning. Analytical modeling and data mining are similar to the idea of machine learning. Here, searching strategies are employed to look for trends and then modify the trial plan as necessary. Those who shop online and receive advice from others related to their purchases are able to identify machine learning. This occurs as commendation enginesutilize ML techniques to find out advertisement releases through the internet in approximately real-time privately.

## II. DEFINITION OF ML

ML is a class of heuristic that is used to provide more accuracy in predicting outcomes with the help of software applications and is not required programmed for that. The basic rule of machine learning is to build algorithms and use statistical analysis to predict an output.

## III. APPLICATION AREA OF MACHINE LEARNING

One of the most recent developments for the digital age that allows for data analysis and prediction is machine learning. In order to improve accuracy and outcomes, industry and researchers use this amazing form of artificial intelligence, which includes extraction, image and medical diagnosis, speech recognition, statistical arbitrage, medical diagnosis, learning associations, learning associations prediction, classification, and regression.

There are several Learning Applications of machine learning that will be elaborated on in the subsequent section.

1. **Image Recognition:** A popular machine learning tool for classifying objects from a batch of data as digital images is called Image Recognition. Digital images, as far as we are aware, are nothing more than combinations of each pixel. The intensity of the image in black and white can be used to calculate the pixel.

2. **Speech Recognition:** Speech recognition techniques are used to convert the verbal words into text and known as “computer speech recognition”, “automatic speech recognition” or “speech to text”.

   Uttered words are recognized by speech recognition software, and machine learning applications translate uttered words into a collection of integers that represent the voice signal. Subsequently, speech signals employ energy intensities in various time-frequency bands to transform them into discrete phonemes or words, enabling unique word recognition.

3. **Medical Diagnosis:** Machine learning encompasses a wide array of methods, techniques, and tools designed to address various challenges within the medical domain, particularly in the realm of analytical and predictive problem-solving. Its applications extend to tasks like identifying and understanding medical parameters, both individually and in combination, to facilitate predictive analyses. For instance, it aids in extracting valuable medical insights

 

for outcomes research, predicting the progression of diseases, and optimizing therapy planning within the context of comprehensive patient management. Furthermore, machine learning plays a crucial role in the analysis of data from Intensive Care Units, enabling the discovery of patterns and trends even in imperfect datasets.

4. **Statistical Arbitrage:** Statistical arbitrage trading policies are being used for temporary analysis and rivet with a huge number of safeties in the analysis. This is used to create a historical correlation among all-purpose financial situations based on amounts that will be taken for more securities. These algorithms can be used for a problem like categorization or opinion or inference difficulty based on the historicalaverage assumption.

5. **Learning Associations:** LA is utilized to establish connections between items and other qualities. This is also helpful in identifying connections between unrelated products and other products based on consumer purchasing patterns.

6. **Classification:** Classification is known as the procedure of insertion of all entity items in a similar group after the study of the population and used independent variables for identification.

For instance, the bank verifies a customer's ability to repay a loan based on factors such as age, income, and past financial history as well as savings.

## IV. MACHINE LEARNING TECHNIQUES

It has been noted that AI significantly enhances the intelligence of computer systems, enabling them to engage in thoughtful processes. This transformation is made feasible through the utilization of machine learning, which facilitates the learning and development of intelligent capabilities within computers. Machine learning encompasses a variety of learning techniques, including supervised, unsupervised, semi-supervised, reinforcement, and evolutionary learning.

1. **Supervised Learning:** Supervised learning (21) relies on having a training dataset that provides guidance for data based on predefined rules. This training process typically yields favorable results when the algorithm is well-suited for the specific problem domain and target. Supervised learning is based on learning from examples and is often used to make predictions with binary outcomes, such as "Yes" or "No" answers. For example, it can answer questions like, "Is it raining today?" or "Does this restaurant meet our quality standards?" with a simple "Yes" or "No." In contrast, regression answers questions involving quantities, such as "How much" or "How many."

2. **Unsupervised Learning:** In this approach, there is no teacher or training set available to provide correct answers. Instead, it seeks to identify similarities between input data by utilizing distance measures and classifies the data based on these similarities. Clustering is a component of supervised learning, wherein the goal is to group similar entities together while distinguishing them from other groups. Various important clustering techniques include the following:

 

- **K-means:** K-means clustering algorithm is an unsupervised algorithm to put K entity into k clusters for n observation with the help of distance measure function. It used the partition concept to put the elements into the cluster using their nearest means.

- **DBSCAN:** Based on the data density, DBSCAN created the group. After that, group the data to identify components in the data point that are comparable. Moreover, classify the low-density area as an outlier and discard it.

- **Hierarchical Clustering:** To create a cluster based on several distance measure functions, such as single linkage distance, complete linkage, and average linkage, hierarchical clustering is employed. They arranged themselves into hierarchical clusters.

3. **Semi-Supervised Learning**: Semi-supervised learning approaches are similar to supervised learning approaches in that they include labeling a small quantity of data and a large amount of unlabeled data, respectively. The concept of semi-supervised learning is positioned amid unsupervised-learning (unlabeled-data) which does not have a training set and supervised learning (labeled-data) which has the training set.

4. **Reinforcement Learning:** This algorithm is grounded in behaviorist psychology and is employed when answers to certain questions are incorrect without specifying the correct approach. It explores multiple possibilities and conducts an analysis to discover the correct answers. Unlike supervised learning, it is akin to a critic who does not offer suggestions for improvement. Reinforcement learning does not deliver precise input and does not provide a predefined output set.

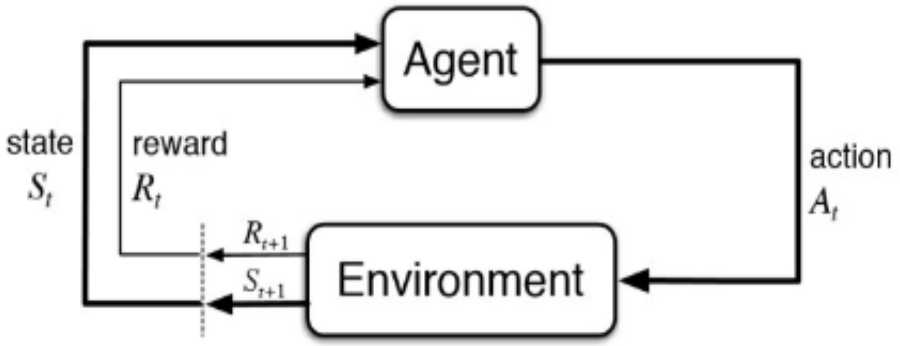


**Figure1.** Diagram for Reinforcement learning (42)

In Figure 1, the agent acquires data from the environment and, based on this data, takes actions. This learning approach differs from Supervised Learning in that there is no predefined answer key for the agent to execute a specific task, whereas in supervised learning, an answer key or training dataset is provided. Figure 2 illustrates the operational concept of Reinforcement learning. Developers devise a mechanism to encourage positive behaviors and penalize unfavorable ones in reinforcement learning. In this method, undesirable behaviors are assigned a negative value, while the agent's desired actions receive a positive value.

 

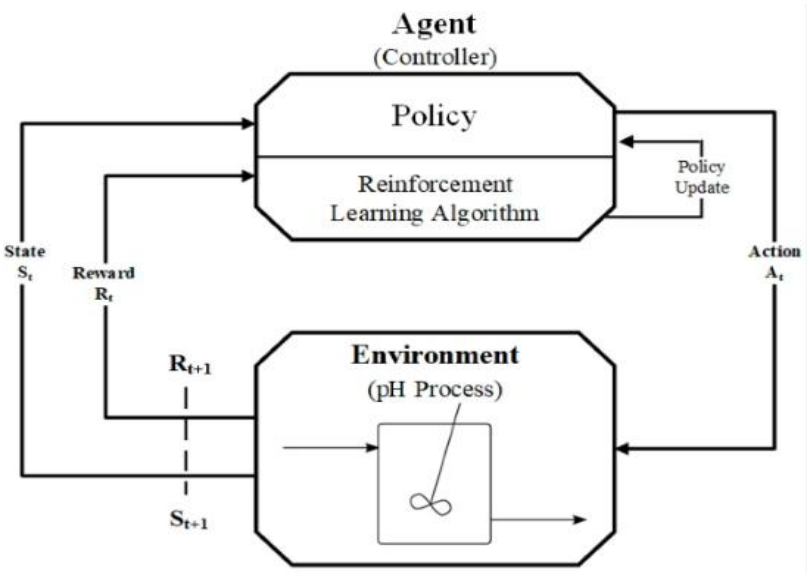


**Figure 2:** Diagram for Reinforcement (43)

5. **Evolutionary Learning:** The foundation of evolutionary learning is biological evolution learning, which may be adjusted in terms of their rates of continuing existence and likelihood of progeny. This one serves as a computer model to verify that the solution is accurate. It has been observed that the Pattern recognition process and data classification are used by Human beings for sensing the environment and analyze the environment. This analysis is always done by the previous environment experience of the selected data for the analysis.

## V. MACHINE LEARNING TECHNIQUES IN DIAGNOSIS AND DETECTION OF DISEASES

After conducting a literature review, it has come to our attention that numerous researchers have expressed the view that a variety of machine learning algorithms are applicable to disease diagnosis. These algorithms have garnered global acceptance due to their high accuracy in diagnosing various diseases. In this chapter of the book, we delve into the application of machine learning techniques for diagnosing diseases such as heart disease, diabetes, and hepatitis.

1. **Heart Disease:** Otoom et al. (5) introduced a system designed to detect and monitor coronary artery disease through an analytical and monitoring process. To conduct their research, the authors utilized a dataset obtained from the UCI machine repository, which included 303 cases with a total of 76 attributes or features. However, out of these 76 attributes, only 13 were selected for use. The authors employed support vector machine, Bayes Net, and Functional Trees (FT) for conducting two tests aimed at heart disease detection. Their analysis was facilitated using the WEKA tool for detection purposes.

 

2. **Diabetes Disease:** A decision tree and a Naive Bayes classifier were used by Iyer et al. (11) to forecast the development of diabetes illness. They linked these illnesses to inadequate insulin production and inappropriate insulin use. The Pima Indian diabetes dataset provided the authors with data, and they used the WEKA data mining tool to perform their tests. Additionally, the authors found that a dataset percentage split of 70% and 30% yielded more accurate cross-validation, with J48 exhibiting 74.86% and 76.95% accuracy. Furthermore, NaiveBayes demonstrated a correctness rate of 79.52% using the Percentage Split method. The authors selected algorithms that achieved the highest success rate and produced more accurate results.

3. **Hepatitis Disease:** Ba-Alwi et al. conducted a comparative analysis of hepatitis disease diagnosis using various algorithms, including FT Tree, J48, Naive Bayes, Naive Bayes updatable, K Star, LMT, and NN. To evaluate the accuracy and execution time of these algorithms, the authors utilized data from the UCI Machine Learning repository. Their comparative analysis included a performance assessment of neural connections and the WEKA algorithm, with neural connections exhibiting lower error rates than WEKA. The authors also employed rough set theory in the analysis of hepatitis disease diagnosis and concluded that it outperformed NN in medical data analysis. Furthermore, they reported that Naive Bayes achieved an accuracy of 96.6% in 0 seconds, while the Naive Bayes Updateable algorithm reached 84% accuracy in the same time frame. The FT Tree algorithm achieved an accuracy of 87.10% in 0 seconds. The authors noted that the Naive Bayes classification algorithm demonstrated high accuracy within a shorter time period.

## VI. MACHINE LEARNING TECHNIQUES IN DISEASE DETECTION

Machine learning (ML) has significantly enhanced modern healthcare by enabling the identification and early detection of critical diseases. It excels at detecting various diseases with higher accuracy and greater speed than physicians. Machine learning, a subset of artificial intelligence (AI), encompasses an array of statistical, probabilistic, and optimization techniques that facilitate learning from historical data and diagnosing diseases based on complex medical records. This technology offers a platform for developing effective heuristics to improve the diagnosis of critical biomedical conditions. Leveraging machine learning can potentially save numerous lives by enabling the earlier and more accurate diagnosis of critical diseases.

## VII. VARIOUS METHODS OF MACHINE LEARNING FOR HEART DISEASES DETECTION

The accuracy of diagnosing different cardiac disorders can be improved by using a variety of machine learning techniques.

1. **Bayes Net, Support Vector Machine (SVM) and Functional Tree (FT**): Otoom et al. introduced a heuristic for diagnosing coronary artery disease. They conducted experiments using three heuristics: Bayes Net, SVM, and FT (5). The Cleveland heart dataset, obtained from the UC Irvine machine learning repository, contains 303 samples and 706 features. However, only 13 out of the 76 features were utilized for diagnosis. The WEKA tool was employed for this purpose. The results indicate that Bayes Net achieved an accuracy of 84.4%, Support Vector Machine achieved 84.9% correctness, and the

 

functional tree provided 83.4% accuracy. SVM is advantageous because it can create accurate classifiers even in noisy environments. However, its primary drawback is that it is a binary classifier and requires pairwise classification for multi-class problems. Additionally, SVM is computationally intensive, making it a relatively slow technique with higher experimental costs.

2. **Hybrid Technique:** Tan et al. (9) proposed a hybrid approach that efficiently combines two machine learning heuristics: Genetic Algorithm (GA) and Support Vector Machine (SVM) using the covering method. For this diagnostic method, data mining tools, namely LIBSVM and WEKA, were employed. Several datasets, including Iris, diabetes-related illnesses, breast cancer, heart conditions, and hepatitis, were chosen from the UCI repository. By applying the hybrid technique that combines GA and SVM, they achieved an accuracy level of 84.07% for the diagnosis of heart diseases.

3. **Naive Bayes, J48, Bagging, and SVM:** Vembandasamy et al. (6) proposed a heuristic for the detection and analysis of heart diseases using the Naive Bayes algorithm. This method leverages Bayes' theorem, making Naive Bayes more effective. The dataset used for this study was obtained from a renowned diabetic research organization in Chennai, containing information about 500 individuals with heart diseases. The analysis was conducted using the Weka tool and classification split. Naive Bayes achieved an accuracy of 86.419%.

   A heuristic for identifying cardiac problems using data mining approaches was described by Chaurasia and Pal (7). They made use of the WEKA tool, which is made up of several machine learning methods, such as Bagging, J48, and Naive Bayes. Just 11 of the 76 attributes in the dataset—which came from the UCI machine learning lab—were taken into account while making the prediction. For Naive Bayes, J48, and Bagging, the corresponding attained accuracies were 81.33%, 83.24%, and 88.12%. On this dataset, bagging fared better in terms of classification accuracy than the other approaches.

   Parthiban et al. (8) developed a machine learning heuristic for detecting and analyzing heart-related illnesses by applying machine learning techniques. They used the WEKA tool to apply the SVM and Naive Bayes algorithms. The results showed that SVM had the highest accuracy, with 74% and 94.60%, respectively, for Naive Bayes and SVM.

   An intelligent computational prediction system was created by Yar Muhammad et al. (44) to aid in the detection and diagnosis of heart illness. This method entailed researching different machine learning classification techniques. Four different feature selection techniques were used to remove noisy and inappropriate features from the dataset in order to improve the quality of the data. The effectiveness of each feature selection technique when used in conjunction with classifiers was evaluated. AUC, F1-score, MCC, ROC curve, accuracy, sensitivity, specificity, and other performance measures were used to assess the efficacy and resilience of the created model. Both the full feature set and the best feature selection were used to evaluate the system's classification accuracy. In addition, P-value and Chi-square tests were performed for each feature selection technique and the ET classifier. It is believed that the technology would be a useful tool to help medical practitioners diagnose heart diseases quickly and

 

accurately.

**Table1:** A Comparative Study to Survey the Various ML Methods for Identification of Various Heart Illnesses.

| ML Techniques | Author | Year | Disease | Resources of Data Set | Tool | Accuracy |
|---|---|---|---|---|---|---|
| Bayes Net | Otoom et al. | 2015 | CAD (coronary artery disease) | UCI | WEKA | 84.5% |
| SVM | Otoom et al. | 2015 | CAD (coronary artery disease) | UCI | WEKA | 85.1% |
| FT | Otoom et al. | 2015 | CAD (coronary artery disease) | UCI | WEKA | 84.5% |
| Naïve Bayes | Vembandasamy et al. | 2015 | Heart Disease | Diabetic Research Institute in Chennai | WEKA | 86.419 % |
| Naïve Bayes | Chaurasia and Pal | 2013 | Heart Disease | UCI | WEKA | 82.31% |
| J48 | Chaurasia and Pal | 2013 | Heart Disease | UCI | WEKA | 84.35% |
| Bagging | Chaurasia and Pal | 2013 | Heart Disease | UCI | WEKA | 85.03% |
| SVM | Parthiban and Srivatsa | 2012 | Heart Disease | Research institute in Chennai | WEKA | 94.60% |
| Naïve Bayes | Parthiban and Srivatsa | 2012 | Heart Disease | Research institute in Chennai | WEKA | 74% |
| Hybrid Technique (GA + SVM) | Tan et al. | 2009 | Heart Disease | UCI | LIBSVM and WEKA | 84.07% |

 

| | | | | | | |
|---|---|---|---|---|---|---|
| Intelligent framework (full features) [Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), Artificial Neural Network (ANN), etc.] | Yar Muhammad et al. | 2021 | Heart Disease | Cleveland heart disease dataset $S_1$ and Hungarian heart disease dataset ($S_2$) | available online at the University of California Irvine (UCI) machine learning repository and UCI Kaggle repository | 92.09% |
| Intelligent framework (selected-features) [Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), Artificial Neural Network (ANN), etc.] | Yar Muhammad et al. | 2021 | Heart Disease | Cleveland heart disease dataset $S_1$ and Hungarian heart disease dataset ($S_2$) | available online at the University of California Irvine (UCI) machine learning repository and UCI Kaggle repository | 94.41% |

A Comparative study to survey the various ML methods for identification of various heart illnesses has been depicted in Table1.

4. **Detection of Diabetic Retinopathy:** Globally, diabetes is an illness that is incredibly common. For the population under 50 years old, it is the most prevalent cause of blindness. Diabetes mellitus is the primary cause of the illness known as diabetic retinopathy. In the retina, it produces microvasculature. If the illness is not treated and

 

worsens, blindness could develop.

Most of the scientists agreed that 85% of these patients can be cured if we can do an early diagnosis. An individual with diabetes is very much susceptible to the threat of diabetic retinopathy (DR) (33). Micro blood vessels are responsible for blood supply to every layer of the retina. These vessels are vulnerable to an uninhibited level of blood sugar. If a large quantity of fructose or glucose gathers in blood, due to inadequate delivery of oxygen to cells, the vessels begin to collapse. Any obstruction in these vessels may lead to a havoc eye injury. And as a consequence, the metabolic rate goes down which leads to structural irregularity in vessels, and that turns to DR (36). Microaneurysms are a prior indication of DR. And this ailment causes changes in the blood vessel's size (enlargement). The symptoms of DR contain microaneurysms (MAs), hemorrhages (HMs) and exudates (EXs), andthe blood vessels' strange enlargement. It takes a lot of time and effort to manually review fundus images to determine whether or not microaneurysms, exudates, blood vessel hemorrhages, and macula have undergone morphological changes. With the use of a computer, this may be completed with ease. Methods for exposing blood vessels in the examination of proliferative diabetic retinopathy are currently being discussed.

At this stage, various indicators of retinopathy are evident, such as microaneurysms (MAs), exudates (EXs), hemorrhages (HMs), and inter-retinal microvascular abnormalities (IRMA). The development of proliferative diabetic retinopathy occurs when abnormal new blood vessels appear in different areas of the retina. This is a complex manifestation of DR that can lead to impaired vision (37). Given the progressive nature of DR, early detection is crucial to preserving a patient's eyesight, emphasizing the need for regular and timely screenings. The implementation of an automated screening system for DR detection could significantly reduce the risk of complete blindness and alleviate the workload of ophthalmologists. Additionally, a computer-aided diagnostic (CAD) system can be designed to distinguish between a retina with potential DR and a healthy one (37–39).

5. **Detection Using Image Processing:** To capture images of the retina, a fundus camera is employed. These images must undergo preprocessing before any subsequent algorithms can be applied.On retinal images, a variety of preprocessing techniques are used, such as adaptive histogram equalization, contrast correction, average filtering, median filtering, and homomorphic filtering. Metrics for performance evaluation, like mean square error (MSE) and peak signal-to-noise ratio (PSNR), are calculated when an algorithm is run on these preprocessed images. A higher PSNR number signifies a better quality processed image.

The two basic kinds of data are those where the disease is present and those where a diagnosis is required. The specificity and sensitivity metrics are used to investigate and evaluate the clinical management's appropriateness.

Numerous methods of detection are employed in various literary works. Among them, 1) region-growing techniques, 2) exudate segmentation, and 3) mathematical morphology techniques are the most often used.

 

6. **Detection Using Machine Learning:** Machine learning techniques can effectively differentiate between EX and non-EX regions, including various types of bright lesions (BLs) like cotton wool spots (CWSs) and drusen. While a classification step was incorporated in several previous studies, we have categorized them as such only when classification was the primary focus.

| Existing Paper | Methodology | Evaluation Parameter |
|---|---|---|
| Tymchenko et al. (48) | Standard CNN | Sensitivity & Specificity |
| Math et al. (49) | Pre-trained CNN | ROC Curve, Sensitivity, Specificity |
| Reddy et al. (50) | An ensemble-based machine learning model comprising of the Machine Learning (ML) Algorithms namely Random Forest classifier, Decision Tree Classifier, Adaboost Classifier, K-Nearest Neighbour classifier, Logistic Regression classifier | Accuracy |

Using the fuzzy c-means clustering technique and a color depiction layout in the Luv color space, Osareh et al. (39) segregated images. García et al. (24) proposed a related method in which the bright image regions were coarsely segmented using a combination of global and adaptive histogram thresholding approaches. Ultimately, a collection of characteristics was taken from every area to assess the effectiveness of three neural network classifiers: Support vector machine (SVM), Multilayer Perceptron (MLP), and Radial basis function (RBF) (24).

Table 2 represents a comparative study of different ML methods for identification of Diabetic Retinopathy.

A multi-scale morphological procedure was suggested by Fleming et al. (38) for the detection of EX. Consequently, depending on their local attributes, SVM was utilized to classify candidate regions as background, drusen, or EX (25).

Table 2. represents a Comparative study to survey the various ML methods for identification of DiabeticRetinopathy

7. **Detection of Hepatitis diseases:** Viral hepatitis is a pressing global public health issue, affecting communities worldwide (17, 18). It is characterized by inflammation of the liver and is primarily caused by viral infections. The estimated global death toll from hepatitis reaches 1.5 million annually. Hepatitis is typically diagnosed through routine blood tests, but the medical diagnosis of this disease is intricate, involving the consideration of

 

various factors. Therefore, automated diagnostic systems can play a vital role in supporting the detection of hepatitis by aiding physicians in making precise decisions.

Various machine learning methods can be employed for the automated diagnosis of this disease. Machine learning encompasses two fundamental learning systems: supervised learning and unsupervised learning. Utilizing machine learning techniques is crucial in the development of decision-making systems for enhanced disease diagnosis.

Clustering algorithms can be applied to enhance classification accuracy. Self-Organizing Map (SOM) has been utilized as a clustering method in several investigations. For classification purposes, various studies have incorporated Support Vector Machine (SVM). Given the significance of disease diagnosis, extensive research is ongoing to devise effective classification methods.

Balkhy et al. (40) presents an innovative machine learning approach for hepatitis disease diagnosis, employing a set of real-world data. Dimensionality reduction is carried out using Non-linear Iterative Partial Least Squares (NIPALS). In summary, the methodology consists of the following steps: 1) Application of SOM for clustering on the experimental dataset, 2) Utilization of NIPALS to reduce dimensionality, enhancing clustering accuracy, and 3) Employing an adaptive Neuro-Fuzzy Inference System (ANFIS) ensemble for hepatitis disease diagnosis.

8. **Hepatitis Disease Detection Using Machine Learning:** It is critical to forecast fatty liver disease (FLD) in order to reduce the risk of major health consequences and to ensure appropriate treatment. This FLD is one of the most prevalent types of liver disease.

Nonalcoholic fatty liver disease (NAFLD) is one of the most prevalent liver diseases. To determine which medical model of NAFLD is the most likely to be predictive, several machine learning methods may be used.

This has become a major public health issue (1, 2). Nonalcoholic steatohepatitis (NASH), fibrosis, and simple steatosis are all included in the spectrum of nonalcoholic fatty liver disease (NAFLD). The common consensus is that uncomplicated steatosis is benign, whereas NASH can lead to cirrhosis, fibrosis, and even hepatocellular cancer (3,4). Moreover, there is a strong correlation between NAFLD and type 2 diabetes, cardiovascular disease, and metabolic problems (11–14). Consequently, early detection is critical since it can greatly improve NAFLD management and prevention.

Liver biopsy represents the conventional diagnostic approach for NAFLD. However, despite its practical accuracy and widespread use in medical diagnosis, ultrasonography falls short in detecting mild steatosis properly (8). Data mining refers to the extraction of specific information and its analysis within large datasets (9, 10). Machine learning (ML) algorithms, which are essentially data-mining tools, come into play here. Machine learning encompasses a range of methods that focus on pattern recognition, including classification and predictive models applied to new data. The four main phases of machine learning are: defining the problem, gathering and preparing data, creating the model, and making the prediction. Eleven modern machine learning approaches (11–15) are available: bagging, AdaBoosting, random forest (RF),

 

aggregating one-dependence estimators (AODE), k-nearest neighbor (kNN), logistic regression (LR), support vector machine (SVM), Bayesian network (BN), naïve Bayes (NB), decision tree (C4.5), and hidden naïve Bayes (HNB).

Preparing and gathering data is the initial step in the machine learning process.

9. **Data Collection and Preprocessing:** BMI was computed using recorded data, including height, weight, and diastolic and systolic blood pressure. These various kinds of data or features will be produced, and if necessary, they will be preprocessed.

## VIII. MACHINE LEARNINGTECHNIQUES

Various machine learning approaches can be utilized to predict this disease. Researchers have the flexibility to implement machine learning algorithms through Python or other programming languages. Some prior work has been conducted using Weka, an open-source software. Weka offers a collection of machine learning algorithms suitable for classification and data mining tasks, making them readily applicable to datasets. Furthermore, Weka features tools for regression, data preprocessing, association rules, clustering, and data visualization, contributing to the advancement of recent machine learning developments.

Data consists of attributes associated with predefined classes or ground truth. Typically, feature selection techniques are employed to choose specific attributes for constructing prediction models. Machine learning methodologies encompass both feature selection and classification. Feature selection aims to eliminate redundant attributes. The primary approach used for developing a suitable prediction model is classification, which can be assessed using metrics such as accuracy, F-measure, precision, recall, and more.

The two main stages of the framework are model construction and prediction. During the model building stage, the goal is to develop a classifier that determines a patient's status based on their medical history and whether or not they have FLD. This study separates features into two categories: advanced features that come from biochemical analyses like blood tests, and basic features that come from standard medical examination results. To extract features based on their information gain scores and other pertinent metrics, a variety of methods are used, including soft computing approaches, redundancy analysis, correlation analysis, the Scott-Knott test, and "out-of-bag" estimates.

## IX. EVALUATION

To evaluate the performance of different FLD prediction algorithms, metrics such as accuracy, precision, specificity, recall (sensitivity), and the F-measure are employed. The assessment considers four potential patient outcomes:

True Positive (TP) - When a patient is predicted to have FLD and indeed has FLD.

False Positive (FP) - When a patient is predicted to have FLD, but, in reality, does not have FLD.

 

False Negative (FN) - When a patient is predicted not to have FLD, but actually has FLD.

True Negative (TN) - When a patient is predicted not to have FLD, and indeed does not have FLD.

These outcomes are used to calculate metrics like accuracy, recall, precision, and the F-measure to assess the algorithm's performance.

Table 3 represents a comparative study of different ML methods for identification of Hepatitis disease.

Table 3. A Comparative study to survey the various ML methods for identification of Hepatitis Disease

| Existing Paper | Methodology | Evaluation Parameter |
|---|---|---|
| Chicco et al. (45) | Ensemble Machine learning | Accuracy |
| Kashif et al. (46) | K Nearest Neighbor, kStar, Naive Bayes, Random Forest, Radial Basis Function, PART, Decision Tree, OneR, Support Vector Machine and Multi-Layer Perceptron" | Accuracy, Recall, Precision, and F-Measure |
| Tian et al. (47) | XGBoost | AUC |

## X. CONCLUSION

The limitations of statistical models, such as their inability to handle categorical data and missing values in extensive datasets, have led to the growing importance of machine learning techniques. Researchers across various domains are employing Machine Learning to explore potential solutions. This chapter aims to present an examination of diverse Machine Learning methods for the detection of different diseases, including diabetes, heart disease, and hepatitis.

A review of prior research reveals that certain algorithms have demonstrated promising results by accurately identifying relevant attributes. In particular, SVM achieves a 94.60% accuracy in heart disease detection, Naive Bayes attains a 95% accuracy in diagnosing diabetes, and the feed-forward neural network effectively classifies hepatitis cases with 98% accuracy. These findings indicate that these algorithms perform well across various disease types, offering insights for potential applications in other domains and suggesting improved methods for future decision-making processes.

 

## REFERENCES

[1] Marsland, S. (2014). *Machine learning: an algorithmic perspective*. Chapman andHall/CRC.
[2] Sharma, P., & Kaur, M. (2013). Classification in pattern recognition: A review. *International Journal of Advanced Research in Computer Science and Software Engineering*,*3*(4).
[3] Rambhajani, M., Deepanker, W., & Pathak, N. (2015). A survey on the implementation of machine learning techniques for dermatology diseases classification. *International Journal of Advances in Engineering & Technology*, *8*(2),194.
[4] Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art, and perspective. *Artificial Intelligence in medicine*, *23*(1),89-109.
[5] Otoom, A. F., Abdallah, E. E., Kilani, Y., Kefaye, A., & Ashour, M. (2015). Effective diagnosis and monitoring of heart disease. *International Journal of Software Engineering and Its Applications*, *9*(1), 143-156.
[6] Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using the Naive Bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, *2*(9),441-444.
[7] Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. *International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol*, *2*,56-66.
[8] Parthiban, G., & Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. *International Journal of Applied Information Systems (IJAIS)*, *3*,2249-0868.
[9] Tan, K. C., Teoh, E. J., Yu, Q., & Goh, K. C. (2009). A hybrid evolutionary algorithm for attribute selection in data mining. *Expert Systems with Applications*, *36*(4),8616-8630.
[10] Karamizadeh, S., Abdullah, S. M., Halimi, M., Shayan, J., & Javad Rajabi, M. (2014, September). Advantages and drawbacks of support vector machine functionality. In *2014 International Conference on Computer, Communications, and Control Technology (I4CT)* (pp. 63-65).IEEE.
[11] Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *arXiv preprintarXiv:1502.03774*.
[12] Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, *3*(2),1797-1801.
[13] Sarwar,A.,&Sharma,V.(2012).IntelligentNaïveBayesapproachtodiagnosediabetesType 2. *International Journal of Computer Applications, IJCA Special Edition Nov*, 14-16.
[14] Ephzibah, E. P. (2011). Cost-effective approach on feature selection using genetic algorithms and fuzzy logic for a diabetes diagnosis. *arXiv preprintarXiv:1103.0087*.
[15] Phyu, T. N. (2009, March). Survey of classification techniques in data mining. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 1, pp.18-20).
[16] Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(01),1.
[17] Ba-Alwi, F. M., & Hintaya, H. M. (2013). Comparative study for analysis the prognostic in hepatitis data: data mining approach. *The spinal cord*, *11*,12.
[18] Karlik, B. (2012). Hepatitis disease diagnosis using backpropagation and the naive bayes classifiers. *IBU Journal of Science and Technology*,*1*(1).
[19] Sathyadevi, G. (2011, June). Application of CART algorithm in hepatitis disease diagnosis. In *2011 International Conference on Recent Trends in Information Technology (ICRTIT)* (pp. 1283-1287).IEEE.
[20] Singh, Y., Bhatia, P. K., & Sangwan, O. (2007). A review of studies on machine learning techniques. *International Journal of Computer Science and Security*, *1*(1),70-84.
[21] https://en.wikipedia.org/wiki/Semi-supervised_learning.
[22] https://searchenterpriseai.techtarget.com/definition/machine-learning-ML.
[23] https://data-flair.training/blogs/machine-learning-applications/
[24] https://towardsdatascience.com/zilic-detect-any-disease-with-machine-learning-fdae88664148.
[25] https://towardsdatascience.com/machine-learning-is-the-future-of-cancer-prediction-e4d28e7e6dfa.
[26] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*.Elsevier.
[27] Mitchell, T. M. (1997). Machinelearning.
[28] Wu, X., & Kumar, V. (Eds.). (2009). *The top ten algorithms in data mining*. CRCpress.
[29] Jiang, L., Zhang, H., & Cai, Z. (2008). A novel Bayes model: Hidden naive Bayes. *IEEE Transactions on knowledge and data engineering*, *21*(10),1361-1371.
[30] Webb, G. I., Boughton, J. R., & Wang, Z. (2005). Not so naive Bayes: aggregating one-dependence estimators. *Machine learning*, *58*(1),5-24.
[31] *Journal of Experimental & Theoretical Artificial Intelligence*, *24*(2),219-230.

 

[32] Yu, L., Jiang, L., Wang, D., & Zhang, L. (2017). Attribute value-weighted average of one- dependence estimators. *Entropy*, *19*(9),501.
[33] Trento, M., Bajardi, M., Borgo, E., Passera, P., Maurino, M., Gibbins, R., & Porta, M. (2002). Perceptions of diabetic retinopathy and screening procedures among diabetic people. *Diabetic Medicine*, *19*(10),810-813.
[34] Singer, D. E., Nathan, D. M., Fogel, H. A., & Schachat, A. P. (1992). Screening for diabetic retinopathy. *Annals of Internal Medicine*, *116*(8),660-671.
[35] Niemeijer, M., Van Ginneken, B., Cree, M. J., Mizutani, A., Quellec, G., Sánchez, C. I., & Wu, X. (2009). Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE transactions on medical imaging*, *29*(1),185-195.
[36] Scotland,G.S.,McNamee,P.,Fleming,A.D.,Goatman,K.A.,Philip,S.,Prescott,G.J.,&Olson,J. A. (2010). Costs and consequences of automated algorithms versus manual grading for the detection of referable diabetic retinopathy. *British Journal of Ophthalmology*, *94*(6), 712-719.
[37] Winder, R. J., Morrow, P. J., McRitchie, I. N., Bailie, J. R., & Hart, P. M. (2009). Algorithms for digital image processing in diabetic retinopathy. *Computerized medical imaging and graphics*, *33*(8), 608-622.
[38] Fleming, A. D., Philip, S., Goatman, K. A., Williams, G. J., Olson, J. A., & Sharp, P. F. (2007). Automated detection of exudates for diabetic retinopathy screening. *Physics in Medicine & Biology*, *52*(24),7385.
[39] Osareh, A., Shadgar, B., & Markham, R. (2009). A computational-intelligence-based approach for detection of exudates in diabetic retinopathy images. *IEEE Transactions on Information Technology in Biomedicine*, *13*(4),535-545.
[40] Balkhy, H. H., El-Saed, A., Sanai, F. M., Alqahtani, M., Alonaizi, M., Niazy, N., & Aljumah, A. (2017). Magnitude and causes of loss to follow-up among patients with viral hepatitis at a tertiary care hospital in Saudi Arabia. *Journal of infection and public health*, *10*(4),379-387.
[41] Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art, and perspective. *Artificial Intelligence in medicine*, *23*(1),89-109.
[42] https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2019/08/agent-environment-reinforcement-learning.png
[43] learning((https://d2h0cx97tjks2p.cloudfront.net/blogs/wp-content/uploads/sites/2/2019/08/agent-environment-reinforcement-learning.png)
[44] Yar Muhammad, Muhammad Tahir, Maqsood Hayat & Kil To Chong, Early and accurate detection and diagnosis of heart disease using intelligent computational model", Article, Open Access, Published: December 2021.
[45] Chicco, Davide, and Giuseppe Jurman. "An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis." *IEEE Access* 9 (2021): 24485-24498.
[46] Kashif, A. A., Bakhtawar, B., Akhtar, A., Akhtar, S., Aziz, N., & Javeid, M. S. (2021). Treatment response prediction in hepatitis C patients using machine learning techniques. *International Journal of Technology, Innovation and Management (IJTIM)*, *1*(2), 79-89.
[47] Tian, X., Chong, Y., Huang, Y., Guo, P., Li, M., Zhang, W., ... & Hao, Y. (2019). Using machine learning algorithms to predict hepatitis B surface antigen seroclearance. *Computational and mathematical methods in medicine*, *2019*.
[48] Tymchenko, Borys, Philip Marchenko, and Dmitry Spodarets. "Deep learning approach to diabetic retinopathy detection." *arXiv preprint arXiv:2003.02261* (2020).
[49] Math, Laxmi, and Ruksar Fatima. "Adaptive machine learning classification for diabetic retinopathy." *Multimedia Tools and Applications* 80.4 (2021): 5173-5186.
[50] Reddy, G. T., Bhattacharya, S., Ramakrishnan, S. S., Chowdhary, C. L., Hakak, S., Kaluri, R.,
[51] & Reddy, M. P. K. (2020, February). An ensemble based machine learning model for diabetic
[52] retinopathy classification. In *2020 international conference on emerging trends in information*
[53] *technology and engineering (ic-ETITE)* (pp. 1-6). IEEE.

 