Scientific Research Publishing

Journal of Intelligent Learning Systems and Applications, 2017, 9, 1-16
http://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

Survey of Machine Learning Algorithms for Disease Diagnostic

Meherwar Fatima[1], Maruf Pasha[2]

[1]Institute of CS & IT, The Women University Multan, Multan, Pakistan
[2]Department of Information Technology, Bahauddin Zakariya University, Multan, Pakistan
Email: maha.bloach@gmail.com, maruf.pasha@bzu.edu.pk

How to cite this paper: Fatima, M. and Pasha, M. (2017) Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, **9**, 1-16.
https://doi.org/10.4236/jilsa.2017.91001

Received: October 17, 2016
Accepted: January 21, 2017
Published: January 24, 2017

Abstract

In medical imaging, Computer Aided Diagnosis (CAD) is a rapidly growing dynamic area of research. In recent years, significant attempts are made for the enhancement of computer aided diagnosis applications because errors in medical diagnostic systems can result in seriously misleading medical treatments. Machine learning is important in Computer Aided Diagnosis. After using an easy equation, objects such as organs may not be indicated accurately. So, pattern recognition fundamentally involves learning from examples. In the field of bio-medical, pattern recognition and machine learning promise the improved accuracy of perception and diagnosis of disease. They also promote the objectivity of decision-making process. For the analysis of high-dimensional and multimodal bio-medical data, machine learning offers a worthy approach for making classy and automatic algorithms. This survey paper provides the comparative analysis of different machine learning algorithms for diagnosis of different diseases such as heart disease, diabetes disease, liver disease, dengue disease and hepatitis disease. It brings attention towards the suite of machine learning algorithms and tools that are used for the analysis of diseases and decision-making process accordingly.

Keywords

Machine Learning, Artificial Intelligence, Machine Learning Techniques

1. Introduction

Artificial Intelligence can enable the computer to think. Computer is made much more intelligent by AI. Machine learning is the subfield of AI study. Various researchers think that without learning, intelligence cannot be developed. There are many types of Machine Learning Techniques that are shown in Figure 1. Supervised, Unsupervised, Semi Supervised, Reinforcement, Evolutionary Learning

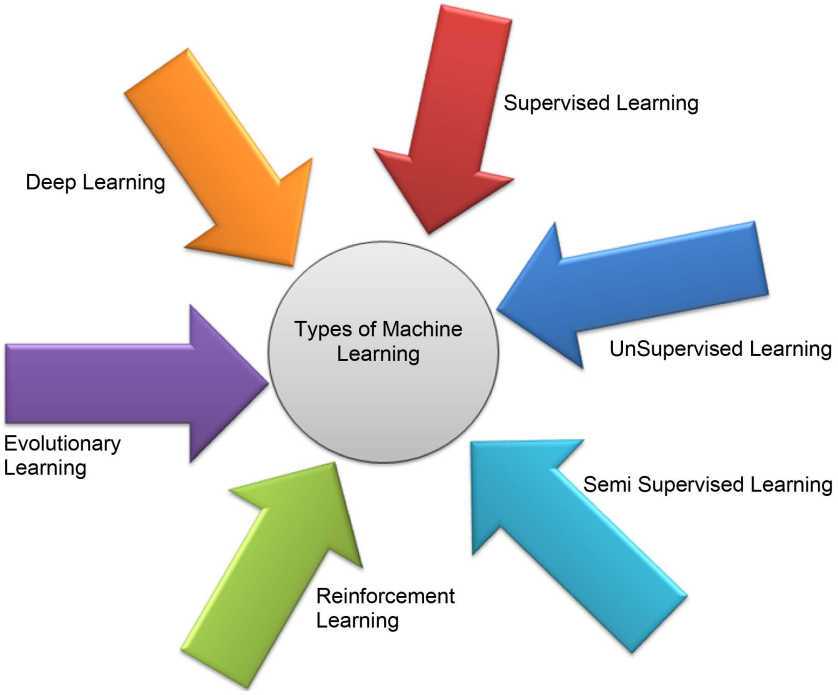

Figure 1. Types of machine learning techniques.

and Deep Learning are the types of machine learning techniques. These techniques are used to classify the data set.

1) Supervised learning: Offered a training set of examples with suitable targets and on the basis of this training set, algorithms respond correctly to all feasible inputs. Learning from exemplars is another name of Supervised Learning. Classification and regression are the types of Supervised Learning.

Classification: It gives the prediction of Yes or No, for example, "Is this tumor cancerous?", "Does this cookie meet our quality standards?"

Regression: It gives the answer of "How much" and "How many".

2) Unsupervised learning: Correct responses or targets are not provided. Unsupervised learning technique tries to find out the similarities between the input data and based on these similarities, un-supervised learning technique classify the data. This is also known as density estimation. Unsupervised learning contains clustering [1].

Clustering: it makes clusters on the basis of similarity.

3) Semi supervised learning: Semi supervised learning technique is a class of supervised learning techniques. This learning also used unlabeled data for training purpose (generally a minimum amount of labeled-data with a huge amount of unlabeled-data). Semi-supervised learning lies between unsupervised-learning (unlabeled-data) and supervised learning (labeled-data).

4) Reinforcement learning: This learning is encouraged by behaviorist psychology. Algorithm is informed when the answer is wrong, but does not inform that how to correct it. It has to explore and test various possibilities until it finds the right answer. It is also known as learning with a critic. It does not recommend improvements. Reinforcement learning is different from supervised learn-

ing in the sense that accurate input and output sets are not offered, nor sub-optimal actions clearly précised. Moreover, it focuses on on-line performance.

5) Evolutionary Learning: This biological evolution learning can be considered as a learning process: biological organisms are adapted to make progress in their survival rates and chance of having off springs. By using the idea of fitness, to check how accurate the solution is, we can use this model in a computer [1].

6) Deep learning: This branch of machine learning is based on set of algorithms. In data, these learning algorithms model high-level abstraction. It uses deep graph with various processing layer, made up of many linear and nonlinear transformation.

Pattern recognition process and data classification are valuable for a long time. Humans have very strong skill for sensing the environment. They take action against what they perceive from environment [2]. Big data turns into Chunks due to multidisciplinary combined effort of machine learning, databases and statistics. Today, in medical sciences disease diagnostic test is a serious task. It is very important to understand the exact diagnosis of patients by clinical examination and assessment. For effective diagnosis and cost effective management, decision support systems that are based upon computer may play a vital role. Health care field generates big data about clinical assessment, report regarding patient, cure, follow-ups, medication etc. It is complex to arrange in a suitable way. Quality of the data organization has been affected due to inappropriate management of the data. Enhancement in the amount of data needs some proper means to extract and process data effectively and efficiently [3]. One of the many machine-learning applications is employed to build such classifier that can divide the data on the basis of their attributes. Data set is divided into two or more than two classes. Such classifiers are used for medical data analysis and disease detection.

Initially, algorithms of ML were designed and employed to observe medical data sets. Today, for efficient analysis of data, ML recommended various tools. Especially in the last few years, digital revolution has offered comparatively low-cost and obtainable means for collection and storage of data. Machines for data collection and examination are placed in new and modern hospitals to make them capable for collection and sharing data in big information systems. Technologies of ML are very effective for the analysis of medical data and great work is done regarding diagnostic problems. Correct diagnostic data are presented as a medical record or reports in modern hospitals or their particular data section. To run an algorithm, correct diagnostic patient record is entered in a computer as an input. Results can be automatically obtained from the previous solved cases. Physicians take assistance from this derived classifier while diagnosing novel patient at high speed and enhanced accuracy. These classifiers can be used to train non-specialists or students to diagnose the problem [4].

In past, ML has offered self-driving cars, speech detection, efficient web search, and improved perception of the human generation. Today machine learning is

present everywhere so that without knowing it, one can possibly use it many times a day. A lot of researchers consider it as the excellent way in moving towards human level. The machine learning techniques discovers electronic health record that generally contains high dimensional patterns and multiple data sets. Pattern recognition is the theme of MLT that offers support to predict and make decisions for diagnosis and to plan treatment. Machine learning algorithms are capable to manage huge number of data, to combine data from dissimilar resources, and to integrate the background information in the study [3].

2. Diagnosis of Diseases by Using Different Machine Learning Algorithms

Many researchers have worked on different machine learning algorithms for disease diagnosis. Researchers have been accepted that machine-learning algorithms work well in diagnosis of different diseases. Figurative approach of diseases diagnosed by Machine Learning Techniques is shown in **Figure 2**. In this survey paper diseases diagnosed by MLT are heart, diabetes, liver, dengue and hepatitis.

2.1. Heart Disease

Otoom *et al.* [5] presented a system for the purpose of analysis and monitoring. Coronary artery disease is detected and monitored by this proposed system. Cleveland heart data set is taken from UCI. This data set consists of 303 cases and 76 attributes/features. 13 features are used out of 76 features. Two tests with three algorithms Bayes Net, Support vector machine, and Functional Trees FT are performed for detection purpose. WEKA tool is used for detection. After

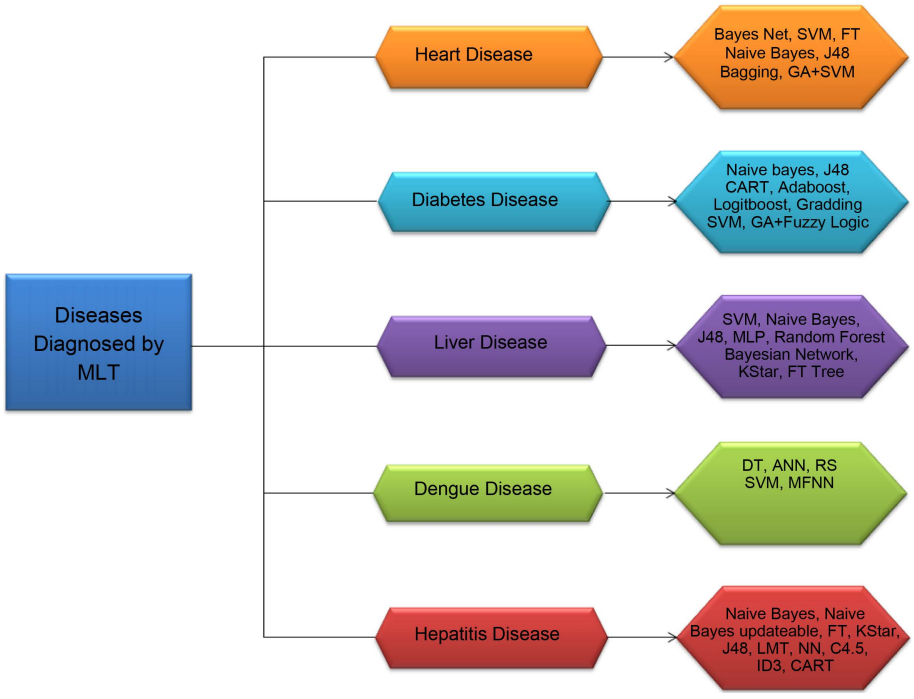

Figure 2. Diseases diagnosed by MLT.

experimenting Holdout test, 88.3% accuracy is attained by using SVM technique. In Cross Validation test, Both SVM and Bayes net provide the accuracy of 83.8%. 81.5% accuracy is attained after using FT. 7 best features are picked up by using Best First selection algorithm. For validation Cross Validation test are used. By applying the test on 7 best selected features, Bayes Net attained 84.5% of correctness, SVM provides 85.1% accuracy and FT classify 84.5% correctly.

Vembandasamy *et al.* [6] performed a work, to diagnose heart disease by using Naive Bayes algorithm. Bayes theorem is used in Naive Bayes. Therefore, Naive Bayes have powerful independence assumption. The employed data-set are obtained from one of the leading diabetic research institute in Chennai. Data set consists of 500 patients. Weka is used as a tool and executes classification by using 70% of Percentage Split. Naive Bayes offers 86.419% of accuracy.

Use of data mining approaches has been suggested by Chaurasia and Pal [7] for heart disease detection. WEKA data mining tool is used that contains a set of machine learning algorithms for mining purpose. Naive Bayes, J48 and bagging are used for this perspective. UCI machine learning laboratory provide heart disease data set that consists of 76 attributes. Only 11 attributes are employed for prediction. Naive bayes provides 82.31% accuracy. J48 gives 84.35% of correctness. 85.03% of accuracy is achieved by Bagging. Bagging offers better classification rate on this data set.

Parthiban and Srivatsa [8] put their effort for diagnosis of heart disease in diabetic patients by using the methods of machine learning. Algorithms of Naive Bayes and SVM are applied by using WEKA. Data set of 500 patients is used that are collected from Research Institute of Chennai. Patients that have the disease are 142 and disease is missing in 358 patients. By using Naive Bayes Algorithm 74% of accuracy is obtained. SVM provide the highest accuracy of 94.60.

Tan *et al.* [9] proposed hybrid technique in which two machine-learning algorithms named Genetic Algorithm (G.A) and Support Vector Machine (SVM) are joined effectively by using wrapper approach. LIBSVM and WEKA data mining tool are used in this analysis. Five data sets (Iris, Diabetes disease, disease of breast Cancer, Heart and Hepatitis disease) are picked up from UC Irvine machine learning repository for this experiment. After applying GA and SVM hybrid approach, 84.07% accuracy is attained for heart disease. For data set of diabetes 78.26% accuracy is achieved. Accuracy for Breast cancer is 76.20%. Correctness of 86.12% is resulting for hepatitis disease. Graphical representation of Accuracy according to time for detection of heart disease is shown in **Figure 3**.

Analysis:

In existing literature, SVM offers highest accuracy of 94.60% in 2012 as in **Table 1**. In many application areas, SVM shows good performance result. Attribute or features used by Parthiban and Srivatsa in 2012 are correctly responded by SVM. In 2015, Otoom *et al.* used SVM variant called SMO. It also uses FS technique to find best features. SVM responds to these features and offers the accuracy of 85.1% but it is comparatively low as in 2012. Training and testing set of both data sets are different, as well as, data types are different.

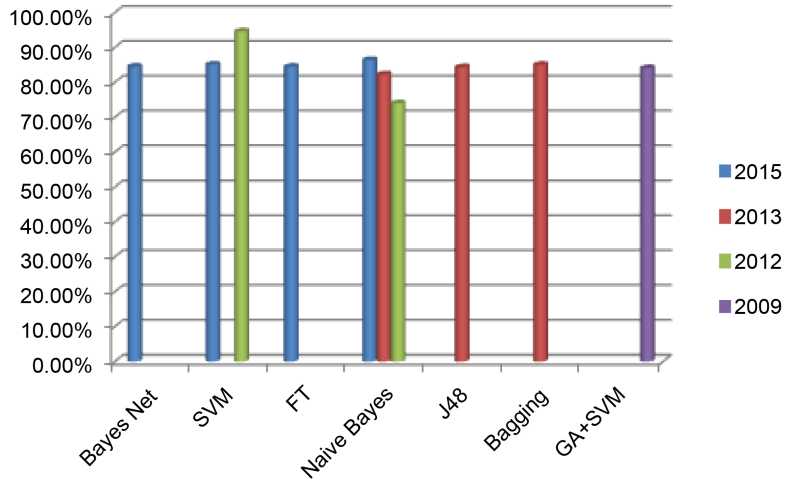

Figure 3. Machine learning algorithm's accuracy to detect heart disease.

Table 1. Comprehensive view of machine learning techniques for heart disease diagnosis.

Machine Learning 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						85.1%
FT						84.5%
Naive Bayes	Vembandasamy *et al.*	2015	Heart Disease	Diabetic Research Institute in Chennai	WEKA	86.419%
Naive Bayes	Chaurasia and Pal	2013	Heart Disease	UCI	WEKA	82.31%
J48						84.35%
Bagging						85.03%
SVM	Parthiban and Srivatsa	2012	Heart disease	Research institute in Chennai	WEKA	94.60%
Naive Bayes						74%
Hybrid Technique (GA + SVM)	Tan *et al.*	2009	Heart disease	UCI	LIBSVM and WEKA	84.07%

Advantages and Disadvantages of SVM:

Advantages: Construct correct classifiers and fewer over fitting, robust to noise.

Disadvantages: It is a binary classifier. For the classification of multi-class, it can use pair wise classification. Its Computational cost is high, so it runs slow [10].

2.2. Diabetes Disease

Iyer *et al.* [11] has performed a work to predict diabetes disease by using decision tree and Naive Bayes. Diseases occur when production of insulin is insufficient or there is improper use of insulin. Data set used in this work is Pima Indian diabetes data set. Various tests were performed using WEKA data mining tool. In this data-set percentage split (70:30) predict better than cross validation. J48 shows 74.8698% and 76.9565% accuracy by using Cross Validation and Percentage Split Respectively. Naive Bayes presents 79.5652% correctness by using PS. Algorithms shows highest accuracy by utilizing percentage split test.

Meta learning algorithms for diabetes disease diagnosis has been discussed by Sen and Dash [12]. The employed data set is Pima Indians diabetes that is received from UCI Machine Learning laboratory. WEKA is used for analysis. CART, Adaboost, Logiboost and grading learning algorithms are used to predict that patient has diabetes or not. Experimental results are compared on the behalf of correct or incorrect classification. CART offers 78.646% accuracy. The Adaboost obtains 77.864% exactness. Logiboost offers the correctness of 77.479%. Grading has correct classification rate of 66.406%. CART offers highest accuracy of 78.646% and misclassification Rate of 21.354%, which is smaller as compared to other techniques.

An experimental work to predict diabetes disease is done by the Kumari and Chitra [13]. Machine learning technique that is used by the scientist in this experiment is SVM. RBF kernel is used in SVM for the purpose of classification. Pima Indian diabetes data set is provided by machine learning laboratory at University of California, Irvine. MATLAB 2010a are used to conduct experiment. SVM offers 78% accuracy.

Sarwar and Sharma [14] have suggested the work on Naive Bayes to predict diabetes Type-2. Diabetes disease has 3 types. First type is Type-1 diabetes, Type-2 diabetes is the second type and third type is gestational diabetes. Type-2 diabetes comes from the growth of Insulin resistance. Data set consists of 415 cases and for purpose of variety; data are gathered from dissimilar sectors of society in India. MATLAB with SQL server is used for development of model. 95% correct prediction is achieved by Naive Bayes.

Ephzibah [15] has constructed a model for diabetes diagnosis. Proposed model joins the GA and fuzzy logic. It is used for the selection of best subset of features and also for the enhancement of classification accuracy. For experiment, dataset is picked up from UCI Machine learning laboratory that has 8 attributes and 769 cases. MATLAB is used for implementation. By using genetic algorithm only three best features/attributes are selected. These three attributes are used by fuzzy logic classifier and provide 87% accuracy. Around 50% cost is less than the original cost. Table 2 provides the Comprehensive view of Machine learning Techniques for diabetes disease diagnosis.

Analysis:

Naive Bayes based system is helpful for diagnosis of Diabetes disease. Naive Bayes offers highest accuracy of 95% in 2012. The results show that this system can do good prediction with minimum error and also this technique is important to diagnose diabetes disease. But in 2015, accuracy offered by Naive Bayes is low. It presents 79.5652% or 79.57% accuracy. This proposed model for detection of Diabetes disease would require more training data for creation and testing. Figure 4 shows the Accuracy graph of Algorithms for the diagnosis of Diabetes disease according to time.

Advantages and Disadvantages of Naive Bayes:

Advantages: It enhances the classification performance by eliminating the unrelated features. Its performance is good. It takes less computational time.

Table 2. Comprehensive view of machine learning techniques for diabetes disease diagnosis.

Machine Learning Techniques	Author	Year	Disease	Resource of Data Set	Tool	Accuracy
Naive Bayes	Iyer *et al.*	2015	Diabetes Disease	Pima Indian Diabetes dataset	WEKA	79.5652%
J48						76.9565%
CART	Sen and Dash	2014	Diabetes Disease	Pima Indian Diabetes dataset from UCI	WEKA	78.646%
Adaboost						77.864%
Logiboost						77.479%
Grading						66.406%
SVM	Kumari and Chitra	2013	Diabetes Disease	UCI	MATLAB 2010a	78%
Naive Bayes	Sarwar and Sharma	2012	Diabetes type-2	Different Sectors of Society in India	MATLAB with SQL Server	95%
GA + Fuzzy Logic	Ephzibah	2011	Diabetes disease	UCI	MATLAB	87%

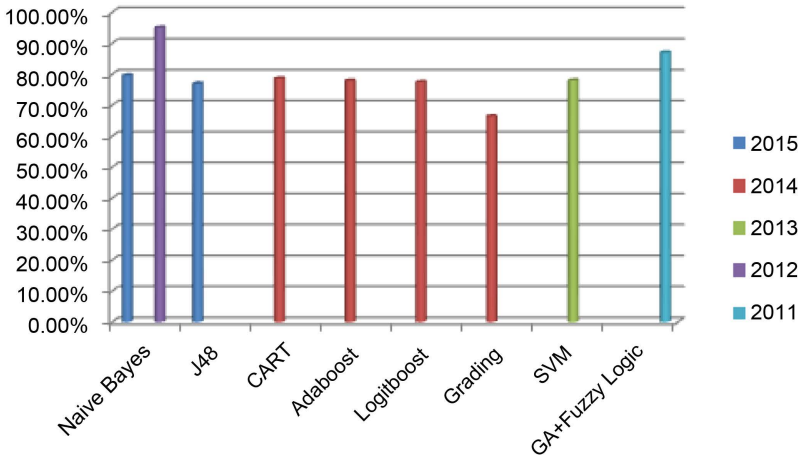

Figure 4. Accuracy of machine learning algorithms to detect diabetes disease.

Disadvantages: This algorithm needs large amount of data to attain good outcomes. It is lazy as they store entire the training examples [16].

2.3. Liver Disease

Vijayarani and Dhayanand [17] predict the liver disease by using Support vector machine and Naive bayes Classification algorithms. ILPD data set is obtained from UCI. Data set comprises of 560 instances and 10 attributes. Comparison is made on the basis of accuracy and time execution. Naive bayes shows 61.28% correctness in 1670.00 ms. 79.66% accuracy is attained in 3210.00 ms by SVM. For implementation, MATLAB is used. SVM shows highest accuracy as compared to the Naive bayes for liver disease prediction. In terms of time execution, Naives bayes takes less time as compared to the SVM.

A study on intelligent techniques to classify the liver patients is performed by the Gulia *et al.* [18]. Used data set is picked up from UCI. WEKA data mining tool and five intelligent techniques J48, MLP, Random Forest, SVM and Bayesian Network classifiers are used in this experiment. In first step, all algorithms are applied on the original data set and get the percentage of correctness. In

second step, feature selection method is applied on whole data-set to get the significant subset of liver patients and all these algorithms are used to test the subset of whole data-set. In third step they take comparison of outcomes before and after feature selection. After FS, algorithms provide highest accuracy as J48 presents 70.669% accuracy, 70.8405% exactness is achieved by the MLP algorithm, SVM provides 71.3551% accuracy, 71.8696% accuracy is offered by Random forest and Bayes Net shows 69.1252% accuracy.

Rajeswari and Reena [19] used the data mining algorithms of Naive Bayes, K star and FT tree to analyze the liver disease. Data set is taken from UCI that comprises of 345 instances and 7 attributes. 10 cross validation test are applied by using WEKA tool. Naive Bayes provide 96.52% Correctness in 0 sec. 97.10% accuracy is achieved by using FT tree in 0.2 sec. K star algorithm classify the instances about 83.47% accurately in 0 sec. On the basis of outcomes, highest classification accuracy is offered by FT tree on liver disease dataset as compared to other data mining algorithms. Table 3 presents the comprehensive view of algorithms for the detection of liver disease.

Analysis:

To diagnose liver disease, FT Tree Algorithm provides the highest result as compare to the other algorithms. When FT tree algorithm is applied on the dataset of liver disease, time taken for result or building the model is fast as compared to other algorithms. According to its attribute, it shows the improved performance. This algorithm fully classified the attributes and offers 97.10% correctness. From the results, this Algorithm plays an important role in determining enhanced classification accuracy of data set. Accuracy graph of algorithms are shown in Figure 5.

Advantages and Disadvantages of FT:

Advantage: Easy to interpret and understand; Fast prediction.

Disadvantage: Calculations are complex mainly if values are uncertain or if several outcomes are linked.

2.4. Dengue Disease

Tarmizi *et al.* [20] performed a work for Malaysia Dengue Outbreak Detection

Table 3. Comprehensive view of machine learning techniques for liver disease diagnosis.

Machine Learning Techniques	Author	Year	Disease	Resource of Data Set	Tool	Accuracy
SVM	Vijayarani and Dhayanand	2015	Liver Disease	ILPD from UCI	MATLAB	79.66%
Naive Bayes						61.28%
J48	Gulia *et al.*	2014	Liver Disease	UCI	WEKA	70.669%
MLP						70.8405%
Random Forest						71.8696%
SVM						71.3551%
Bayesian Network						69.1252%
Naive Bayes	Rajeswari and Reena	2010	Liver Disease	UCI	WEKA	96.52%
K Star						83.47%
FT tree						97.10%

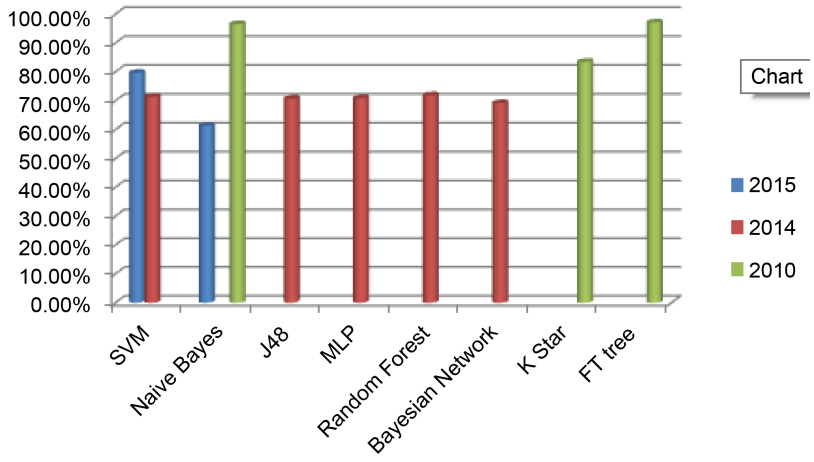

Figure 5. Accuracy of machine learning algorithms to detect liver disease.

by using the Models of Data Mining. Dengue is becoming a severe contagious disease. It creates trouble in those countries where weather is humid for example Thailand, Indonesia and Malaysia. Decision Tree (DT), Artificial Neural Network (ANN), and Rough Set Theory (RS) are the classification algorithms that are used in this study to predict dengue disease. Data set are taken from Public Health Department of Selangor State. WEKA data mining tool with two tests (10 Cross-fold Validation and Percentage split) is used. By using 10-Cross fold validation DT offers 99.95% accuracy, ANN presents 99.98% of Correctness and RS shows 100% accuracy. After using PS, Both Decision tree and Artificial Neural Network gives 99.92% of correctness. RS achieves 99.72% accuracy.

Fathima and Manimeglai [21] performed a work to predict Arbovirus-Dengue disease. Data mining algorithm that are used by these researchers are Support Vector Machine. Data set for analysis is obtained from King Institute of Preventive Medicine and surveys of many hospitals and laboratories of Chennai and Tirunelveli from India. It contains 29 attributes and 5000 samples. Data is examined by R project version 2.12.2. Accuracy that is achieved by SVM is 0.9042.

Ibrahim *et al.* [22] suggested a system in which Artificial neural network is used for forecasting the defervescence day of fever in patients of dengue disease. Only clinical signs and symptoms are used by the proposed system for detection. The data are gathered from 252 hospitalized patients, in which 4 patients are having DF (Dengue fever) and 248 patients are having DHF (dengue hemorrhagic fever). MATLAB's neural network toolbox is used. Algorithm of Multilayer feed-forward neural network (MFNN) is used in this experiment. Day of defervescence of fever is accurately predicted by MFNN in DF and DHF with 90% correctness.

Figure 6 shows the accuracy graph of all algorithms for the diagnosis of Dengue disease.

Analysis:

Different Machine learning techniques are used to diagnose dengue disease. Dengue disease is one of the serious contagious diseases. As in Table 4, for detection of dengue disease, RS theory shows the highest result as compared to the

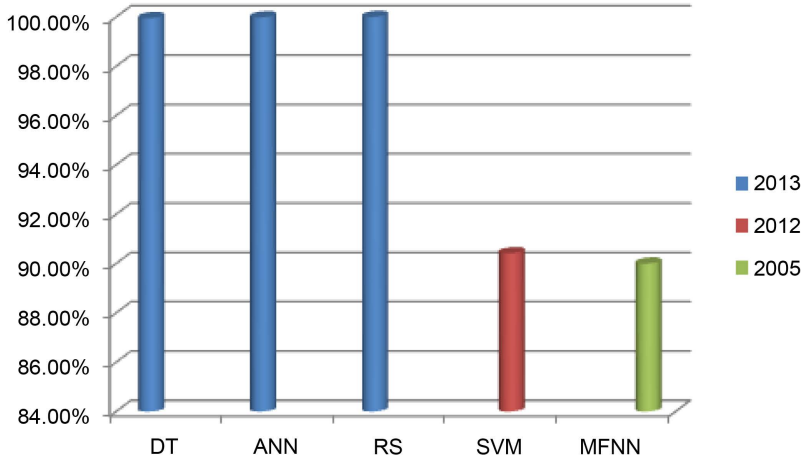

Figure 6. Accuracy of machine learning algorithms for dengue disease.

Table 4. Analysis of machine learning techniques for dengue disease detection.

Machine Learning Techniques	Author	Year	Disease	Resource of Data Set	Tool	Accuracy
DT						99.95%
ANN	Tarmizi *et al.*	2013	Dengue Disease	Public Health Department of Selangor State	WEKA	99.98%
RS						100%
SVM	Fathima and Manimeglai	2012	Arbovirus-Dengue disease	King Institute of Preventive Medicine and surveys of many hospitals and laboratories of Chennai and Tirunelveli from India	R project Version 2.12.2	90.42%
MFNN	Ibrahim *et al.*	2005	Dengue disease	From 252 hospitalized patients	MATLAB neural network Tool box	90%

other algorithms. In 2005 and 2012, researchers used different algorithms but did not attain highest result and improvements. In 2013, accuracy is improved by using RS. It is capable to manage uncertainty, noise and missing data. For the purpose of classification, Developed RS classifier is based on the Rough set theory. Selection of attribute empowers the classifier to surpass the other models. RS is a promising rule based method that offers meaningful information. RS is also best from neural network in term of time. NN takes much time to build model. DT is complex as well as costly algorithm. RS does not need any initial and additional information about data but Decision tree needs information.

Advantages and Disadvantages of RS:

Advantages: It is very easy to understand and provides direct understanding of attained result. It evaluates data significance. It is appropriate for both qualitative and quantitative data. It discovers the hidden patterns. It also finds minimal set of data. It can find relationship that cannot be identified by statistical methods.

Disadvantages: It has not so many limitations still it is not widely used.

2.5. Hepatitis Disease

Ba-Alwi and Hintaya [23] suggested a comparative analysis. Data mining algorithms that are used for hepatitis disease diagnosis are Naive Bayes, Naive Bayes

updatable, FT Tree, K Star, J48, LMT, and NN. Hepatitis disease data set was taken from UCI Machine Learning repository. Classification results are measured in terms of accuracy and time. Comparative Analysis is taken by using neural connections and WEKA: data mining tool. Results that are taken by using neural connection are low than the algorithms used in WEKA. In this Analysis of Hepatitis disease diagnosis, second technique that is used is rough set theory, by using WEKA. Performance of Rough set procedure is better than NN specially in case of medical data analysis. Naive Bayes gives the accuracy of 96.52% in 0 sec. 84% Accuracy is attained by the Naive Bayes Updateable algorithm in 0 sec. In 0.2 sec FT Tree presents the accuracy of 87.10%. K star offers 83.47% Correctness. Time taken for K star algorithm is 0 sec. Correctness of 83% is achieved by J48 and time that J48 takes to classify is 0.03 sec. LMT provides 83.6% accuracy 0.6 sec. Neural network shows 70.41% of correctness. Naive Bayes is best classification algorithm used in the rough set technique. It offers high accuracy in minimum time.

Karlik [24] shows a comparative analysis of Naive Bayes and back propagation classifiers to diagnose hepatitis disease. Key advantage of using these classifiers is that they require small amount of data for categorization. Types of hepatitis are "A, B, C, D and E". These are generated by different viruses of hepatitis. Rapid Miner open source software is used in this analysis. Hepatitis data set is taken from UCI. Data set include 20 features and 155 instances. 15 attributes are used in this experiment. Naive Bayes classifier gives 97% accuracy. Three-layered feed-forward NN are used and trained with Back propagation algorithm 155 instances are used for training. Correctness of 98% is attained.

Sathyadevi [25] employed C4.5, ID3 and CART algorithms for diagnosing the disease of hepatitis. This study uses the UCI hepatitis patient data set. WEKA, tool is used in this analysis. CART has offered great performance handling of missing values. So, CART algorithm shows a highest classification accuracy of 83.2%. ID3 Algorithm offers 64.8% of accuracy. 71.4% is attained by C4.5 algorithm. Binary decision tree (DT) that is generated by CART algorithm has only two or no child. DT that is formed by the C4.5 and ID3 can have two or more children. CART algorithm performs well in terms of Accuracy and time complexity.

Analysis:

Many algorithms have been used for diagnosis of different diseases. **Table 5** gives the comprehensive view. For the detection of Hepatitis disease, Feed forward neural network with back propagation shows highest accuracy of 98%. Because in this model, three layered feed forward neural network is trained with error back propagation algorithm. Back propagation training with the rule of delta learning is an iterative gradient algorithm planned to lessen the RMSE "root mean square error" between the real output of a multilayered feed-forward neural networks and a desired output. Every layer is connected to preceding layer and having no other connection. Second best result is offered by Naive Bayes. But in terms of time to build model, Naive Bayes runs fast as compare to neural network. Fi-

gurative approach for the detection of hepatitis is shown in Figure 7.

Advantages and Disadvantages of NN:

Advantages: Adaptive Learning, Self-Organization, Real Time Operation Fault Tolerance via Redundant Information Coding.

Disadvantages: Less over fitting needs great computational effort. Sample Size must be large. It's time consuming. Engineering Judgment does not develop the relations between input and output variables so that the model behaves like a black box [26].

3. Discussions and Analysis of Machine Learning Techniques

For diagnosis of Heart, Diabetes, Liver, Dengue and Hepatitis diseases, several machine-learning algorithms perform very well. From existing literature, it is observed that Naive Bayes Algorithm and SVM are widely used algorithms for

Table 5. Comprehensive view of machine learning techniques for hepatitis disease.

Machine Learning Techniques	Author	Year	Disease	Resource of Data Set	Tool	Accuracy
Naive Bayes	Ba-Alwi and Hintaya,	2013	Hepatitis Disease	UCI	WEKA	96.52%
Naive Bayes updateable						84%
FT						87.10%
K Star						83.47%
J48						83%
LMT						83.6%
NN						70.41%
Naive Bayes	Karlik	2011	Hepatitis Disease	UCI	Rapid Miner	97%
Feed forward NN with Back propagation						98%
C4.5						71.4%
ID3	Sathyadevi	2011	Hepatitis Disease	UCI	WEKA	64.8%
CART						83.2%

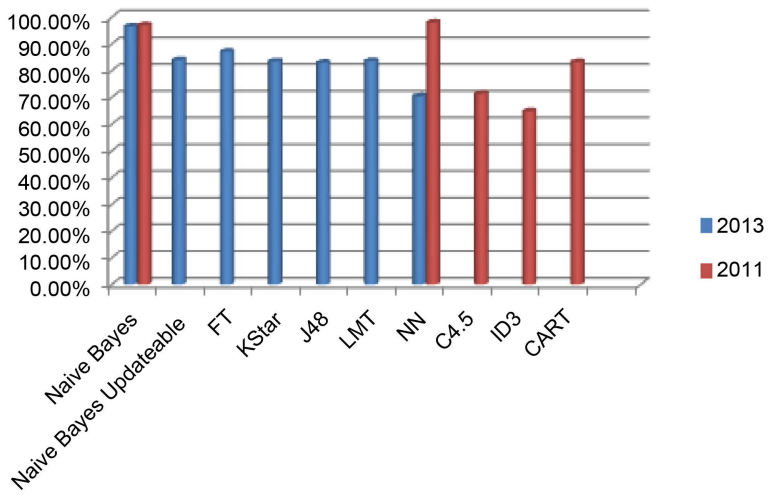

Figure 7. Machine learning algorithm's accuracy to detect hepatitis disease.

detection of diseases. Both algorithms offer the better accuracy as compare to other algorithms. Artificial Neural network is also very useful for prediction. It also shows the maximum output but it takes more time as compared to other algorithms. Trees algorithm are also used but they did not attain wide acceptance due to its complexity. They also shows enhanced accuracy when it responded correctly to the attributes of data set. RS theory is not widely used but it presents maximum output.

4. Conclusion

Statistical models for estimation that are not capable to produce good performance results have flooded the assessment area. Statistical models are unsuccessful to hold categorical data, deal with missing values and large data points. All these reasons arise the importance of MLT. ML plays a vital role in many applications, e.g. image detection, data mining, natural language processing, and disease diagnostics. In all these domains, ML offers possible solutions. This paper provides the survey of different machine learning techniques for diagnosis of different diseases such as heart disease, diabetes disease, liver disease, dengue and hepatitis disease. Many algorithms have shown good results because they identify the attribute accurately. From previous study, it is observed that for the detection of heart disease, SVM provides improved accuracy of 94.60%. Diabetes disease is accurately diagnosed by Naive Bayes. It offers the highest classification accuracy of 95%. FT provides 97.10% of correctness for the liver disease diagnosis. For dengue disease detection, 100% accuracy is achieved by RS theory. The feed forward neural network correctly classifies hepatitis disease as it provides 98% accuracy. Survey highlights the advantages and disadvantages of these algorithms. Improvement graphs of machine learning algorithms for prediction of diseases are presented in detail. From analysis, it can be clearly observed that these algorithms provide enhanced accuracy on different diseases. This survey paper also provides a suite of tools that are developed in community of AI. These tools are very useful for the analysis of such problems and also provide opportunity for the improved decision making process.

References

[1] Marshland, S. (2009) Machine Learning an Algorithmic Perspective. CRC Press, New Zealand, 6-7.

[2] Sharma, P. and Kaur, M. (2013) Classification in Pattern Recognition: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*, **3**, 298.

[3] Rambhajani, M., Deepanker, W. and Pathak, N. (2015) A Survey on Implementation of Machine Learning Techniques for Dermatology Diseases Classification. *International Journal of Advances in Engineering & Technology*, **8**, 194-195.

[4] Kononenko, I. (2001) Machine Learning for Medical Diagnosis: History, State of the Art and Perspective. *Journal of Artificial Intelligence in Medicine*, **1**, 89-109.

[5] Otoom, A.F., Abdallah, E.E., Kilani, Y., Kefaye, A. and Ashour, M. (2015) Effective Diagnosis and Monitoring of Heart Disease. *International Journal of Software En-*

gineering and Its Applications. **9**, 143-156.

[6] Vembandasamy, K., Sasipriya, R. and Deepa, E. (2015) Heart Diseases Detection Using Naive Bayes Algorithm. *IJISET-International Journal of Innovative Science, Engineering & Technology*, **2**, 441-444.

[7] Chaurasia, V. and Pal, S. (2013) Data Mining Approach to Detect Heart Disease. *International Journal of Advanced Computer Science and Information Technology* (*IJACSIT*), **2**, 56-66.

[8] Parthiban, G. and Srivatsa, S.K. (2012) Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients. *International Journal of Applied Information Systems* (*IJAIS*), **3**, 25-30.

[9] Tan, K.C., Teoh, E.J., Yu, Q. and Goh, K.C. (2009) A Hybrid Evolutionary Algorithm for Attribute Selection in Data Mining. *Journal of Expert System with Applications*, **36**, 8616-8630. https://doi.org/10.1016/j.eswa.2008.10.013

[10] Karamizadeh, S., Abdullah, S.M., Halimi, M., Shayan, J. and Rajabi, M.J. (2014) Advantage and Drawback of Support Vector Machine Functionality. 2014 *IEEE International Conference on Computer, Communication and Control Technology* (*I4CT*), Langkawi, 2-4 September 2014, 64-65. https://doi.org/10.1109/i4ct.2014.6914146

[11] Iyer, A., Jeyalatha, S. and Sumbaly, R. (2015) Diagnosis of Diabetes Using Classification Mining Techniques. *International Journal of Data Mining & Knowledge Management Process* (*IJDKP*), **5**, 1-14. https://doi.org/10.5121/ijdkp.2015.5101

[12] Sen, S.K. and Dash, S. (2014) Application of Meta Learning Algorithms for the Prediction of Diabetes Disease. *International Journal of Advance Research in Computer Science and Management Studies*, **2**, 396-401.

[13] Kumari, V.A. and Chitra, R. (2013) Classification of Diabetes Disease Using Support Vector Machine. *International Journal of Engineering Research and Applications* (*IJERA*), **3**, 1797-1801.

[14] Sarwar, A. and Sharma, V. (2012) Intelligent Naïve Bayes Approach to Diagnose Diabetes Type-2. *Special Issue of International Journal of Computer Applications* (0975-8887) *on Issues and Challenges in Networking, Intelligence and Computing Technologies-ICNICT* 2012, **3**, 14-16.

[15] Ephzibah, E.P. (2011) Cost Effective Approach on Feature Selection using Genetic Algorithms and Fuzzy Logic for Diabetes Diagnosis. *International Journal on Soft Computing* (*IJSC*), **2**, 1-10. https://doi.org/10.5121/ijsc.2011.2101

[16] Archana, S. and DR Elangovan, K. (2014) Survey of Classification Techniques in Data Mining. *International Journal of Computer Science and Mobile Applications*, **2**, 65-71

[17] Vijayarani, S. and Dhayanand, S. (2015) Liver Disease Prediction using SVM and Naïve Bayes Algorithms. *International Journal of Science, Engineering and Technology Research* (*IJSETR*), **4**, 816-820.

[18] Gulia, A., Vohra, R. and Rani, P. (2014) Liver Patient Classification Using Intelligent Techniques. (*IJCSIT*) *International Journal of Computer Science and Information Technologies*, **5**, 5110-5115.

[19] Rajeswari, P. and Reena,G.S. (2010) Analysis of Liver Disorder Using Data Mining Algorithm. *Global Journal of Computer Science and Technology*, **10**, 48-52.

[20] Tarmizi, N.D.A., Jamaluddin, F., Abu Bakar, A., Othman, Z.A., Zainudin, S. and Hamdan, A.R. (2013) Malaysia Dengue Outbreak Detection Using Data Mining Models. *Journal of Next Generation Information Technology* (*JNIT*), **4**, 96-107.

[21] Fathima, A.S. and Manimeglai, D. (2012) Predictive Analysis for the Arbovirus-

Dengue using SVM Classification. *International Journal of Engineering and Technology*, **2**, 521-527.

[22] Ibrahim, F., Taib, M.N., Abas, W.A.B.W., Guan, C.C. and Sulaiman, S. (2005) A Novel Dengue Fever (DF) and Dengue Haemorrhagic Fever (DHF) Analysis Using Artificial Neural Network (ANN). *Computer Methods and Programs in Biomedicine*, **79**, 273-281. https://doi.org/10.1016/j.cmpb.2005.04.002

[23] Ba-Alwi, F.M. and Hintaya, H.M. (2013) Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach. *International Journal of Scientific & Engineering Research*, **4**, 680-685.

[24] Karlik, B. (2011) Hepatitis Disease Diagnosis Using Back Propagation and the Naive Bayes Classifiers. *Journal of Science and Technology*, **1**, 49-62.

[25] Sathyadevi, G. (2011) Application of CART Algorithm in Hepatitis Disease Diagnosis. *IEEE International Conference on Recent Trends in Information Technology* (*ICRTIT*), MIT, Anna University, Chennai, 3-5 June 2011, 1283-1287.

[26] Singh, Y., Bhatia, P.K., and Sangwan, O. (2007) A Review of Studies on Machine Learning Techniques. *International Journal of Computer Science and Security*, **1**, 70-84.