Grenze International Journal of Engineering and Technology, June Issue

Performance Evaluation of Various Approaches for Disease Prediction using Weka

Sushil Kalmegh[1], Prerna S. Tayade[2] and Amol P. Bhagat[3]
[1]PG Department of Computer Science, Sant Gadge Baba Amravati University, Amravati (M.S.) 444 602, India
Email: peru.tayade@gmail.com
[2]Department of Information Technology, Prof Ram Meghe College of Engineering & Management, Badnera, Amravati (M.S.) 444 701, India
Email: amol.bhagat84@gmail.com

***Abstract*—This paper introduces Weka, which uses data mining algorithms to classify viruses. This paper introduces Weka's proposed data mining algorithm for classifying viruses. This paper presents a review and evaluation of a data mining algorithm for classifying diseases using Weka. The data mining algorithms ZeroR, Linear Regression, Multilayer Perceptron, Random Forest, Simple K-Means, Hierarchical Clustering, and Farthest First are proposed for disease classification using Weka. Weka acts as the judge in evaluating these algorithms, which possess the capability to accurately classify any ailment. Their performance is scrutinized based on both precision and the time taken to classify diseases.**

***Index Terms*— Data mining, data preprocessing, classification, Weka tool, disease prediction.**

I. INTRODUCTION

Electronic health records (EHRs) are collected, integrated, and analyzed to create appropriate treatment plans and treatment strategies for patients. Electronic medical records make it easier to diagnose, diagnose and treat diseases. However, existing EHR models are critical. Many problems such as big data, privacy and private information limit the development of centralized medical information. A novel approach and algorithm [1] have been introduced for handling the extraction of medical data that is distributed across various locations (such as hospitals and clinics) within an organization. However, due to organizational rules, this medical information cannot be transferred to an external site. Hence, the demand for global computing necessitates segmentation into local computing units to facilitate information dissemination across the network. The capacity to decompose computations should be sufficiently versatile to address diverse distributions and participants within each computing instance. Notably, every data source is depicted by an agent. The agent then deals with international names or exchanges some small orders with other agents or goes anywhere and does the usual work that can be done in any local location. The goal is to perform international operations with minimal communication or travel by participants in the network, in a manner that preserves the confidentiality and security of local information. The rule of thumb is used to predict heart disease using only actual heart disease data. This information resides in different medical facilities and cannot be moved to a central location.

Recently, privacy methods have been developed for horizontal and vertical distributed systems based on the star topology required by the agent [2]. Each site distributes information through an untrusted third-party intermediary. This tool is responsible for creating integrated forecast models. Nevertheless, these models incur

Grenze ID: 01.GIJET.10.2.501_1

significant communication expenses because of frequent data exchanges with central agents. That's why, a new predictive model [1] was developed to collect data from multiple referral sources that do not transmit confidential patient data. Therefore, the patient's privacy is protected. A decentralized algorithm has been added to my organization rules based on trust rules and weighted rules based on big local data. Rather than devising individual rules for each local user, the integrated model generalizes the extracted rules and stores them in distributed storage independently.

Each treatment center's information is based on a different region (hospital or hospital). Figure 1 shows an EHR model with a complete (multiple local facilities such as a pharmacy, hospital, and clinic) and a hybrid model. Each nursery is located in a different area. Moving data from one local location to another is not a good solution for many reasons, including data ownership, privacy, communication overhead, and big data. However, building decision models based on medical data requires big data.

Information is used only through the extraction and decision making of many important confidential information in many healthcare applications [3]. The research in [4] offers a thorough examination of the data mining process within the medical domain. It furnishes descriptive insights into various aspects, including the nature of data, data storage, business intelligence, and data mining techniques. Data mining tools use patients' original medical records, test history, and personal information. Hence, when symptoms are identified, proactive measures can be taken to mitigate or prevent the onset of diseases, thereby enabling doctors to minimize their impact. Utilizing data mining in healthcare yields numerous benefits, including the detection of health insurance fraud, disease prediction and diagnosis, hospital bed management, and the development of personalized treatment plans based on recommended medications for individual patients [5].

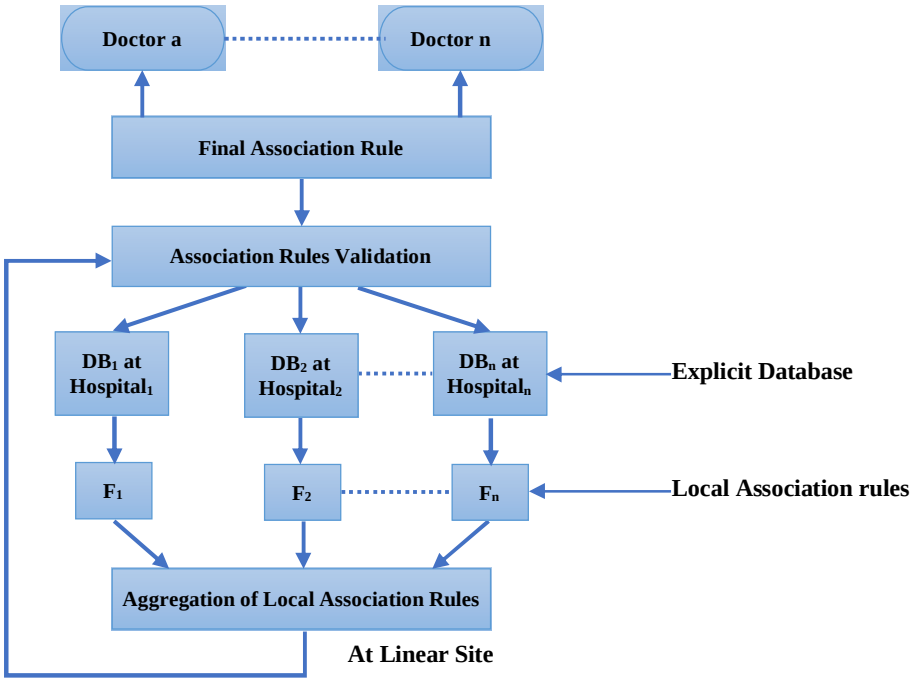


Figure 1. EHRs Model [1]

In [6], the authors discuss a research study focusing on information retrieval and tools within the medical domain. They explore the utilization of various data analysis tools, such as classification, clustering, correlation, and regression. For instance, neural networks (ANN) have been employed to forecast skin cancer, and diverse ANN parameters were utilized in [7] to predict the likelihood of non-melanoma tumors. Additionally, convolutional neural networks (CNN) can predict disease progression in disease populations and diagnose cerebral infarction [8]. Despite their benefits, neural networks require high processing times and are difficult to interpret [9]. Predicting chronic kidney disease based on environmental health data using classification techniques such as nearest neighbor (KNN). In contrast to neural networks (NN), KNN is straightforward to employ and train, although it's susceptible to noise. Another classification method, the decision tree (DT), is known for its interpretability and ability to handle both numerical and categorical data. However, it's constrained

to a single output and its effectiveness varies depending on the dataset type, rendering its performance unstable. Various decision tree models [10] have been applied to predict liver diseases.

The performance evaluations of various decision tree models, including J48, Logical Model Trees (LMT), Random Forests, Random Trees, Reduced Error Pruned Trees (REPTree), Decision Logs, and Hoeffding Trees, are detailed. Other research endeavors explore different classification approaches, such as the fusion of decision trees with neural networks, aimed at enhancing prediction and classification [11, 12, 13]. These studies leverage decision trees (DT) and neural networks (NN) for predicting lung disease and heart disease, respectively.

Support Vector Machine (SVM) emerges as a prominent tool in medical research due to its superior accuracy compared to alternative methods. SVM classifiers encounter fewer issues and can effectively handle heterogeneous data. Numerous health studies employ SVM, such as the utilization of genetic SVM technology for heart disease identification. This model identifies crucial features and analyzes output signals from heart valve ultrasound scans. In [14], SVM was employed to assess breast cancer and classify breast cancer cases.

However, it's noted that SVM can be costly as it necessitates precise data and generates distinct models for each dataset.

Association rules (AR) play a crucial role in medical data mining as they explore intriguing patterns and connections between diseases and symptoms. Their application has expanded to include the development of predictive models [14]. In particular, association rules are utilized to identify correlations between brain perfusion measurements in different regions as an early diagnostic tool for Alzheimer's disease. The initial step involves predicting neurons within the region of interest, which serve as input for the association rule algorithm. The association between brain region ratios and support and confidence measures is then determined. Using common rules from different products can achieve better and better results compared to distributors. The performance of each classifier and the performance of each combination are discussed separately. In [15], researchers introduced a memory-efficient and rapid association rule mining algorithm built upon the MapReduce framework. The work in [16] defines a new concept of multiple rule generation. Given operational constraints and privacy concerns, the aforementioned activities rely on sequential methods, which are not well-suited for managing distributed medical data. The mission to improve the performance of healthcare organizations has spurred research into the integration of distributed health information. Various data integration systems have been added to approximately 900 databases.

Recognizing patients with immunoglobulin-resistant antibodies is critical for prompt treatment and diagnosis of Kawasaki disease, highlighting the necessity for efficient risk assessment tools. Data-driven methodologies can identify high-risk individuals by discerning intricate patterns within real-time data. To enable clinical prediction of immunoglobulin resistance and address challenges stemming from missing clinical data and non-interpretability of machine learning models, a multi-tiered approach was devised. This approach integrates the extraction of missing data patterns with pattern comprehension [17]. First, by combining clinical and patient characteristics, the common method is used to characterize missing data to complete the group according to selection and missing information; Prediction models tailored to specific patient subgroups are crafted based on the available data. Subsequently, group lasso is employed for specialized selection to identify exceptional cases. Third, the Explanation Augmentation Machine is a study defined as an additive model and used to predict entire patient groups. The fusion of integrated and supervised learning techniques has tackled the limitations of medical data mining and advocated for a data-driven approach to predictive learning and action, thereby enhancing clinical decision-making [17].

Kawasaki disease (KD) manifests as a vascular ailment commonly observed in infants and children, with potential to cause severe cardiac complications. It has emerged as a leading cause of pediatric heart disease globally, with its incidence on the rise in recent times. Despite ongoing research, the exact etiology and pathophysiology of KD remain largely elusive. Around 25% of KD patients develop persistent aneurysms, with these aneurysms accounting for 5% of heart-related issues in adults under 40 [18]. Hence, timely diagnosis and treatment are imperative for favorable outcomes.

In clinical settings, prompt administration of intravenous immunoglobulin (IVIG) and aspirin therapy upon diagnosis may mitigate stroke risks by up to 5%. However, approximately 15–20% of KD patients exhibit resistance to IVIG therapy, elevating their cardiovascular disease risk [19]. Recent studies suggest that tailored primary care adjusted for individual risk factors may enhance outcomes for KD patients. Consequently, early prediction of IVIG resistance holds significant promise in aiding physicians' decision-making, emphasizing the need for effective risk assessment tools.

Exploring Electronic Health Records (EHRs) to compile comprehensive medical data encompassing patient diagnoses, medication history, and participation in clinical trials holds promise for advancing medical research and fostering groundbreaking discoveries. Leveraging data analytics presents an invaluable opportunity to

uncover hidden patterns and develop predictive models aimed at enhancing health outcomes across self-diagnosis, diagnosis, and clinical intervention [20].
A data-centric approach has been adopted to pinpoint patients resistant to intravenous immunoglobulin (IVIG) treatment, facilitating the utilization of EHR data and computational methodologies to deliver timely and efficacious care. Notably, numerous clinical investigations have been undertaken to explore the clinical and laboratory factors influencing the necessity for additional treatment in IVIG-resistant patients with CH.
The efficacy of six models employing different risk factors for predicting IVIG resistance was assessed in a cohort of 504 KD patients [6]. Eight independent risk factors were identified, and a predictive model was devised using multiple regression analysis [21]. Another study employed multivariate logistic regression analysis to forecast non-response to IVIG, concluding that numerous factors associated with IVIG resistance exhibited poor predictive performance, as evidenced by low AUC scores [22]. To reconcile conflicting findings across relevant publications, a meta-analysis was conducted to delineate predictors of IVIG resistance [19]. Furthermore, in addition to traditional statistical methods, machine learning techniques have emerged as valuable tools for risk assessment. Data from 767 KD patients, including 170 initially IVIG-resistant individuals, were collected [23]. Random forest analysis was applied to identify IVIG resistance among KD patients [23].

II. ZeroR

The ZeroR is the simplest method for classification, based on the target and ignoring all different predictions. The ZeroR classifier only predicts most groups (classes). Although ZeroR is not predictive, it is useful in determining performance as a benchmark for other classification methods. Creates frequency tables of targets and selects frequency values. The ZeroR model for the data set where golf = yes and line is 0.64 is shown in Figure 2. Since ZeroR does not use any of these, there is nothing to predict contributions to the model. ZeroR can only predict most of the classes correctly. As mentioned before, ZeroR is only suitable for determining the performance of alternative distributions.

Figure 2. ZeroR Approach

III. Linear Regression

Prediction using linear regression is described in more detail. Linear regression provides a straightforward equation format for estimating outcomes based on given input variables. To illustrate, consider the task of predicting weight (y) based on height (x). The linear regression model for this scenario can be expressed as follows:

$$y = B_0 + B_1 * x_1$$

or

$$weight = B_0 + B_1 * height$$

Where B_0 is the difference coefficient and B_1 is the height coefficient. Use learning techniques to find optimal coefficients. Once you find it, enter the height difference to estimate the weight. For example, let's use $B_0 = 0.1$ and $B_1 = 0.5$. Let's attach the weight of a person 182 cm tall and calculate it in kilograms

$$weight = 0.1 + 0.5 * 182$$
$$weight = 91.1$$

The above equation can be plotted as two straight lines as shown in Figure 3. Regardless of the height, B_0 is the threshold. Run from different heights from 100 to 250 centimetres and substitute the equation and get the weight value to form a straight line.

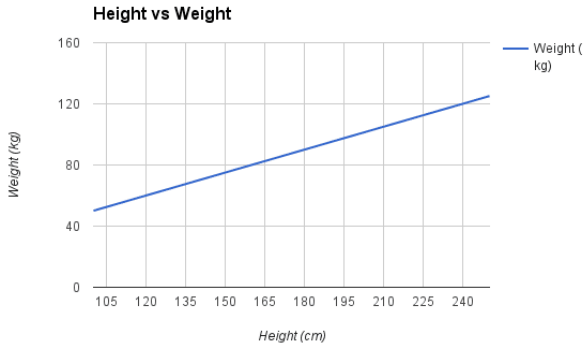


Figure 3. Linear Regression

IV. Multilayer Perceptron

Multilayer perceptrons fall into the category of feedforward algorithms in that the components are combined with the initial weight in the weight and are affected by activation just like perceptron's. However, the distinction lies in how each combination line extends to the subsequent layer. Each layer furnishes the following layer with computed results, serving as an internal representation of the data. This path goes from hidden layer to output layer.

However, there's more to it. If the algorithm merely computes the weight in each neuron, propagates the outcome to the output layer, and halts, it won't learn the weights necessary for minimizing the cost. Without multiple iterations, genuine learning cannot occur. This is where backpropagation comes into play. Backpropagation is a learning technique enabling multiple layers within the network to adjust weights, thereby minimizing processing costs, as depicted in Figure 4. Backpropagation necessitates certain conditions to function effectively. Functions used to combine inputs and weights in neurons (such as numerical weights) and threshold functions (like ReLU) must differ. Moreover, these functions require bounded derivatives, as gradient descent is the prevalent optimization technique utilized in multilayer perceptrons.

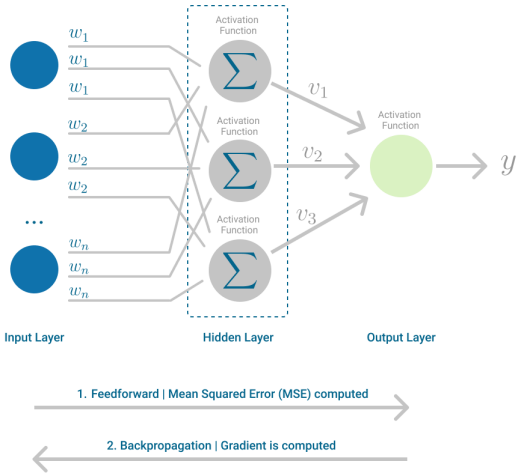


Figure 4. Multilayer Perceptron the Feedfoward and Backpropagation Steps

During each iteration, the gradient of the mean square error is computed for all input-output pairs and propagated through all weight layers. Subsequently, the weights of the first hidden layer are adjusted based on the gradient value to refine the process. This mechanism allows the weights to be updated across the neural network. Mathematically, one iteration of gradient descent can be represented as:

$$\Delta_{\omega}(t) = -\varepsilon \frac{dE}{d\omega(t)} + \alpha \Delta_{\omega}(t-1)$$

$\Delta_{\omega}(t)$: Gradient Current Iteration
ε: Bias
E: Error
$\omega(t)$: Weight Vector
α: *Learning Rate*
$\Delta_{\omega}(t-1)$: Gradient Previous Iteration

This process persists until the gradient of each input-output pair converges. Convergence occurs when the newly calculated gradient changes less than the specified convergence threshold compared to the previous iteration.

V. RANDOM FOREST

ERandom Forest stands as a prevalent supervised machine learning algorithm applied extensively in classification and regression tasks. It constructs a decision tree ensemble comprising various models and conducts majority voting on the distribution and mean in regression scenarios. Notably, a key advantage of the random forest algorithm lies in its ability to handle datasets containing both continuous variables, suitable for regression, and categorical variables, adept for distribution. It particularly excels in classification problems.
The algorithm is detailed below as shown in Figure 5.

- **Step 1:** In a random forest, n random samples are selected from a dataset comprising k data points.
- **Step 2:** Create a separate decision tree for each model.
- **Step 3:** Every tree decides to reproduce.
- **Step 4:** Final results are determined by the majority of votes or the distribution average and returned accordingly.

Hyperparameters play a vital role in enhancing the performance of the random forest model, aiding in gauging its efficacy and accelerating its computational speed. Applying hyperparameters increases predictive power:

- **n_estimators**: The quantity of trees produced by the algorithm prior to aggregating predictions through averaging.
- **max_features:** Random forest defines the highest count of features available for node splitting within the distribution.
- **mini_sample_leaf:** Establish the minimum leaf count required for internal separation.

Following hyper-parameters surges the speed of the random forest.

- **n_jobs:** This parameter specifies the number of processors permitted for computation. When set to 1, only one processor is utilized, while setting it to -1 imposes no restrictions.
- **random_state:** This parameter governs the randomness within the model. If the model maintains consistent state values alongside identical hyperparameters and training data, it ensures reproducibility, yielding the same results across executions.

oob_score: OOB, short for "out of bag," denotes a cross-validation technique employed in random forests. In this approach, data not used during training is utilized for performance evaluation in the model's third phase. These models are referred to as out-of-bag models.

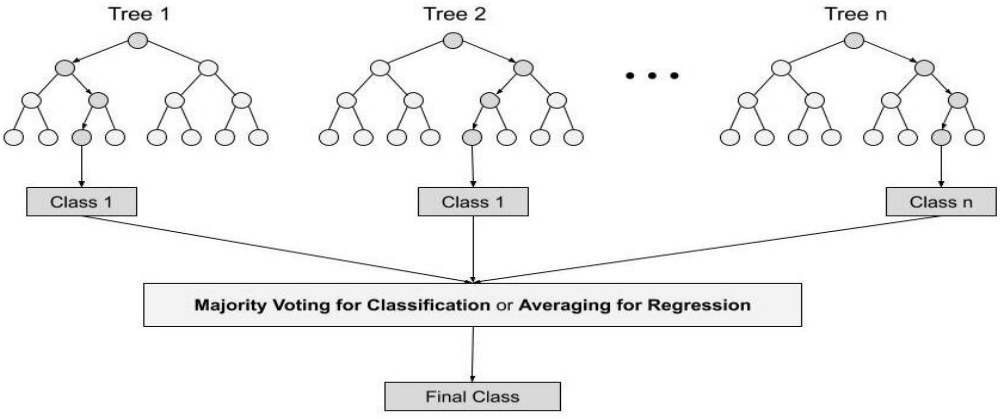


Figure 5. Random Forest Algorithm

VI. Simple K-Means

The technique utilized by k-means to address the problem is known as expectation maximization. In the Expectation (E) step, data points are assigned to their nearest cluster, while in the Maximization (M) step, the mean of each cluster is computed:

$$J = \sum_{i=1}^{m}\sum_{k=1}^{K} \omega_{ik}||x^i - \mu_k||^2$$

where ω_{ik}=1 for data point x^i if it belongs to cluster k; otherwise, ω_{ik}= 0. Also, μ_k is the centroid of x^i's cluster.

It's a minimization problem of two parts. First minimize J with respect to ω_{ik} and treat μ_k fixed. Then minimize J with respect to μ_k and treat ω_{ik} fixed. Technically speaking, differentiate J with respect to ω_{ik} first and update cluster assignments (E-step). Then differentiate J with respect to μ_k and re-compute the centroids after the cluster assignments from previous step (M-step). Therefore, E-step is:

$$\frac{\partial J}{\partial \omega_{ik}} = \sum_{i=1}^{m}\sum_{k=1}^{K} \omega_{ik}||x^i - \mu_k||^2$$

$$\Rightarrow \omega_{ik} = \begin{cases} 1 \; if k = argmin_j ||x^i - \mu_j||^2 \\ 0 \qquad otherwise \end{cases}$$

In other words, assign the data point $x^i$to the closest cluster judged by its sum of squared distance from cluster's centroid. M-step is:

$$\frac{\partial J}{\partial \mu_k} = 2\sum_{i=1}^{m} \omega_{ik}(x^i - \mu_k) = 0$$

$$\Rightarrow \omega_k = \frac{\sum_{i=1}^{m} \omega_{ik} x^i}{\sum_{i=1}^{m} \omega_{ik}}$$

This entails the recalculation of the centroid for each cluster to accommodate the updated position. As clustering algorithms, such as k-means, rely on distance metrics to gauge data point similarity, it is advisable for these distances to fall within zero to 1 standard deviation of the dataset. This range is often preferred given that dataset features typically adhere to this standardization. It's quantified in diverse units, like age and income, reflecting the varied nature of the data. Due to the inherent characteristics of k-means and the potential for inadequate centroid initialization at the algorithm's onset, disparate initializations may result in distinct clusters. This variance arises because the k-means algorithm might become trapped in local optima, failing to converge towards the global optimum. Consequently, it's advisable to execute the algorithm using various starting points for centroids and select the run yielding the minimal squared distance. Remaining consistent with sampling equates to maintaining the same group allocation.

$$\frac{1}{m_k}\sum_{i=1}^{m_k} ||x^i - \mu_c^k||^2$$

VII. Hierarchical Classification/ Clustering

Agglomerative Clustering holds significant relevance in the business realm, making it a central aspect of this research endeavor. Once the agglomerative clustering type is understood, divisive hierarchical clustering becomes straightforward. Below is the algorithm for conducting hierarchical classification/clustering:

Step 1: Initially, allocate all points to individual classes/clusters.

Distinct colors are utilized to denote separate classes/clusters. In this instance, there are five distinct classes/clusters corresponding to the five data points.

Step 2: Proceed by examining the smallest distance within the proximity matrix and combining the points exhibiting the smallest distance. Subsequently, update the proximity matrix.

ID	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

In this scenario, the smallest distance measures 3, leading to the merging of points 1 and 2.

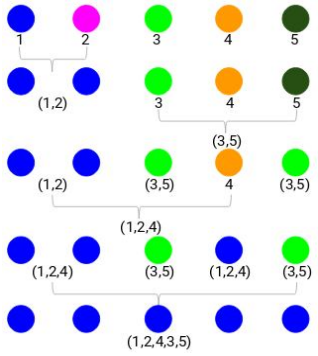
The updated classes/ clusters and the proximity matrix

Student_ID	Marks
(1,2)	10
3	28
4	20
5	35

In this context, the maximum value between two tokens (7, 10) is selected to update the category/token. This process can involve determining the highest, lowest, or average price. Subsequently, recalibrate the proximity matrix for the groups/clusters.

ID	(1,2)	3	4	5
(1,2)	0	18	10	25
3	18	0	8	7
4	10	8	0	15
5	25	7	15	0

Step 3: Continue iterating through step 2 until only a solitary class/cluster remains.
Initially, scan the proximity matrix for the minimum distance, then merge the nearest class/group pair accordingly. After repeating these steps, combined groups/groups are obtained as shown below evident, we commenced with 5 distinct classes/clusters and eventually converged into a single class/cluster. This encapsulates the operational principle of agglomerative hierarchical classification/clustering.



VIII. Farthest First

Farthest First is a variant of K; this means giving each location the center of the location furthest from a current location. The subject must be in the data field. Overall this results in a much faster deployment because less space is required and fewer updates are required.

IX. Evaluation of Data Mining Algorithms for Disease Classification

The proposed data mining algorithms namely ZeroR, Linear Regression, Multilayer Perceptron, Random Forest, Simple K-Means, Hierarchical Clustering, and Farthest First used for disease classification are described in the previous section. This section provides the evaluation of these algorithms. The evaluation is carried out using Weka 3.8.6 on Windows 10 with 4 GB RAM, 1TB HDD and Intel i3 8th generation processor. The comparison of ZeroR, Linear Regression, Multilayer Perceptron, Random Forest, Simple K-Means, Hierarchical Clustering, and Farthest First on the basis of accuracy and classification time is presented in table 1. It can be observed that Farthest First is the most computationally efficient algorithm compared to remaining evaluated algorithms. Among the remaining assessed algorithms, Simple K-Means demonstrates the highest level of accuracy. The Hierarchical Clustering has the worst accuracy. While Multilayer Perceptron has more computational complexity.

Table I. Comparative Analysis of Data Mining Algorithms for Disease Classification

Algorithm	Accuracy	Classification Time (Seconds)
ZeroR	93.72%	0.23
Linear Regression	78%	0.94
Multilayer Perceptron	89.24%	3.06
Random Forest	86.98%	1.96
Simple K-Means	94.03%	0.08
Hierarchical Clustering	48.52%	0.62
Farthest First	91.82%	0.02

X. Conclusion

In this paper data mining approaches for disease classification are presented. The Weka for disease classification using data mining algorithms is described in this chapter. The data pre-processing and algorithm evaluation using Weka is described in this chapter. The data mining algorithms ZeroR, Linear Regression, Multilayer Perceptron, Random Forest, Simple K-Means, Hierarchical Clustering, and Farthest First are proposed for disease classification using Weka. The algorithms are evaluated using Weka. These algorithms have the capability to classify a wide range of diseases. These algorithms are compared on the basis of accuracy and disease classification time. The data mining algorithms ranging from highest to lowest has accuracy of Simple K-Means (94.03%), ZeroR (93.72%), Farthest First (91.82%), Multilayer Perceptron (89.24%), Random Forest (86.98%), Linear Regression (78%), and Hierarchical Clustering (48.52%). If arranged on the basis of the computational complexity lowest to highest disease classification time for Farthest First (0.02), Simple K-Means (0.08), ZeroR (0.23), Hierarchical Clustering (0.62), Linear Regression (0.94), Random Forest (1.96), and Multilayer Perceptron (3.06). These algorithms can be further evaluated for more diseases.

References

[1] [Khedr et al., 2021] AHMED M. KHEDR, ZAHER AL AGHBARI, AMAL AL ALI, AND MARIAM ELJAMIL, “An Efficient Association Rule Mining From Distributed Medical Databases for Predicting Heart Diseases,” IEEE Access, VOLUME 9, pp. 15320- 15333, Jan 2021.

[2] [Khedr et al., 2018] A. Khedr, Z. A. L. Aghbari, and I. Kamel, ``Privacy preserving decomposable mining association rules on distributed data," Int. J. Eng. Technol., vol. 7, nos. 313, pp. 157162, 2018.

[3] [Cifci and Hussain, 2018] M. A. Cifci and S. Hussain, ``Data mining usage and applications in health services," Int. J. Informat. Vis., vol. 2, no. 4, p. 225, Jun. 2018, doi: 10.30630/joiv.2.4.148.

[4] [Ray, 2018] R. Ray, ``Advances in data mining: Healthcare applications," Int. Res. J. Eng. Technol., vol. 5, no. 3, pp. 23562395, Mar-2018

[5] [Nayak, 2017] P. Nayak, ``A survey on medical data by using data mining techniques," Int. J. Advance Res., Ideas Innov. Technol., vol. 3, no. 6, pp. 13301335, 2017.

[6] [Tomar and Agarwal, 2013] D. Tomar and S. Agarwal, ``A survey on data mining approaches for healthcare," Int. J. Bio-Sci. Bio-Technol., vol. 5, no. 5, pp. 241266, Oct. 2013.

[7] [Roffman et al., 2018] D. Roffman, G. Hart, M. Girardi, C. J. Ko, and J. Deng, ``Predicting nonmelanoma skin cancer via a multi-parameterized articial neural network," Sci. Rep., vol. 8, no. 1, p. 1701, Dec. 2018.

[8] [Chen et al., 2017] 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, pp. 88698879, 2017.
[9] [Subhashini and Jeyakumar, 2017] R. Subhashini and M. K. Jeyakumar, ``OF-KNN technique: An approach for chronic kidney disease prediction," Int. J. Pure Appl. Math., vol. 116, no. 24, pp. 331348, 2017.
[10] [Nahar and Ara, 2018] N. Nahar and F. Ara, ``Liver disease prediction by using different decision tree techniques," Int. J. Data Mining Knowl. Manage. Process, vol. 8, no. 2, pp. 19, Mar. 2018.
[11] [Mathan et al., 2018] K. Mathan, P. M. Kumar, P. Panchatcharam, G. Manogaran, and R. Varadharajan, ``A novel Gini index decision tree data mining method with neural network classiers for prediction of heart disease," Des. Automat. Embedded Syst., vol. 22, no. 3, p. 225, 2018.
[12] [Hsu et al., 2018] C.-H. Hsu, G. Manogaran, P. Panchatcharam, and S. Vivekanandan, ``A new approach for prediction of lung carcinoma using back propagation neural network with decision tree classiers," in Proc. IEEE 8th Int. Symp. Cloud Service Comput. (SC), Nov. 2018, pp. 111115.
[13] [Vidic et al., 2018] I. Vidi¢, L. Egnell, N. P. Jerome, J. R. Teruel, T. E. Sjøbakk, A. Østlie, H. E. Fjøsne, T. F. Bathen, and P. E. Goa, ``Support vector machine for breast cancer classication using diffusion-weighted MRI histogram features: Preliminary study: Machine learning in DWI of breast cancer," J. Magn. Reson. Imag., vol. 47, no. 5, pp. 12051216, May 2018.
[14] [Ramasamy and Nirmala, 2020] S. Ramasamy and K. Nirmala, ``Disease prediction in data mining using association rule mining and keyword based clustering algorithms," Int. J. Comput. Appl., vol. 42, no. 1, pp. 18, Jan. 2020.
[15] [Chengyan et al., 2020] L. I. Chengyan, S. Feng, and G. Sun, ``DCE-miner: An association rule mining algorithm for multimedia based on the MapReduce framework," Multimedia Tools Appl., vol. 79, pp. 123, Jun. 2020.
[16] [Taer et al., 2020] P. Y. Ta³er, K. U. Birant, and D. Birant, ``Multitask-based association rule mining," Turkish J. Electr. Eng. Comput. Sci., vol. 28, no. 2, pp. 933955, Mar. 2020.
[17] [Wang et al., 2020] HAOLIN WANG, ZHILIN HUANG, DANFENG ZHANG, JOHAN ARIEF, TIEWEI LYU, AND JIE TIAN, "Integrating Co-Clustering and Interpretable Machine Learning for the Prediction of Intravenous Immunoglobulin Resistance in Kawasaki Disease," IEEE Access, VOLUME 8, pp. 97064- 97071, June 2020.
[18] [McCrindle et al., 2017] B. W. McCrindle, A. H. Rowley, J. W. Newburger, J. C. Burns, A. F. Bolger, M. Gewitz, A. L. Baker, M. A. Jackson, M. Takahashi, P. B. Shah, T. Kobayashi, M.-H.Wu, T. T. Saji, and E. Pahl, ``Diagnosis, treatment, and long-term management of Kawasaki disease: A scientic statement for health professionals from the American Heart Association," Circulation, vol. 135, no. 17, pp. e927-e999, 2017
[19] [Li et al., 2018] X. Li, Y. Chen, Y. Tang, Y. Ding, Q. Xu, L. Sun, W. Qian, G. Qian, L. Qin, and H. Lv, ``Predictors of intravenous immunoglobulin-resistant kawasaki disease in children: A meta-analysis of 4442 cases," Eur. J. Pediatrics, vol. 177, no. 8, pp. 12791292, Aug. 2018.
[20] [Collins and Moons, 2019] G. S. Collins and K. G. M. Moons, ``Reporting of articial intelligence prediction models," Lancet, vol. 393, no. 10181, pp. 15771579, 2019.
[21] [Tan et al., 2019] X.-H. Tan, X.-W. Zhang, X.-Y. Wang, X.-Q. He, C. Fan, T.-W. Lyu, and J. Tian, ``A new model for predicting intravenous immunoglobin-resistant kawasaki disease in chongqing: A retrospective study on 5277 patients," Sci. Rep., vol. 9, no. 1, p. 1722, Dec. 2019.
[22] [Kim et al., 2018] M. K. Kim, M. S. Song, and G. B. Kim, ``Factors predicting resistance to intravenous immunoglobulin treatment and coronary artery lesion in patients with Kawasaki disease: Analysis of the Korean nationwide multicenter survey from 2012 to 2014," Korean Circulat. J., vol. 48, no. 1, pp. 7179, 2018.
[23] [Takeuchi et al., 2017] M. Takeuchi, R. Inuzuka, T. Hayashi, T. Shindo, Y. Hirata, N. Shimizu, J. Inatomi, Y. Yokoyama, Y. Namai, Y. Oda, M. Takamizawa, J. Kagawa, Y. Harita, and A. Oka, ``Novel risk assessment tool for immunoglobulin resistance in kawasaki disease: Application using a random forest classier," Pediatric Infectious Disease J., vol. 36, no. 9, pp. 821826, Sep. 2017.