# An Efficient Lymph Disease Prediction Model Using Naïve Bayes Tree Classifier

**[1]K. Bhuvaneswari**
**[1]Assistant Professor, Department of Computer Application, Idhaya College for Women Kumbakonam**
Email: bonish88@gmail.com

**Abstract**

This research expects to build up a model to upgrade lymphatic diseases analysis by the utilization of irregular forest group machine-learning method prepared with a basic examining plan. This examination has been completed in two noteworthy stages: highlight choice and characterization. In this study, Naïve Bayes tree (NBtree) is employed to classify the data in Lymph disease identification. Exploratory outcomes show that the proposed method accomplishes an amazing improvement in characterization precision rate.

**Keywords:** Classification, Lymph; NB; Decision trees

## 1. Introduction

In these days, Computer-Aided Diagnosis (CAD) applications have turned out to be one of the key research themes in medical biometrics diagnostic tasks. Medical diagnosis relies on the experience of the doctor adjacent to the current data. Therefore, various articles recommended a few methodologies to process the doctor's investigation and judgment tasks about genuine clinical evaluations [1]. With sensible achievement, machine-learning systems have been connected in developing the CAD applications because of its solid ability of removing complex connections in the medical data [2]. Crude medical data requires some powerful characterization procedures to help the computer-based examination of such voluminous and heterogeneous data. Precision of clinically analyzed cases is especially critical issue to be considered amid order. Much of the time, the measure of medical datasets is typically incredible, which straightforwardly influences the intricacy of the data mining technique [3]. In this way, the huge scale medical data is viewed as a wellspring of huge difficulties in data mining applications, which includes extricating the most spellbinding or discriminative highlights. In this way, include decrease has a noteworthy job in wiping out unimportant highlights from medical datasets [4], [5]. Dimensionality decrease technique expects to diminish computational unpredictability with the conceivable focal points of upgrading the general arrangement execution. It incorporates dispensing with unimportant

highlights before model execution, which makes screening tests quicker, progressively viable and less exorbitant and this is an essential prerequisite in medical applications [6].

The lymphatic system is a fundamental piece of the invulnerable system in expelling the interstitial liquid from tissues. It retains and transports fats and fat-dissolvable nutrients from the digestive system and conveys these supplements to the cells of the body. It transports white blood cells to and from the lymph nodes into the bones. In addition, it transports antigen-displaying cells to the lymph nodes where a safe reaction is invigorated. Distinctive medical imaging procedures have been utilized for the examination of the lymphatic channels and lymph glands status [7]. The present condition of lymph nodes with got data from lymphography method can learn the grouping of the examined diagnosis [8]. The extension of lymph nodes can be a list to a few conditions and stretches out to increasingly noteworthy conditions that risk life [9]. The investigation of the lymph nodes is critical in diagnosis, prognosis, and treatment of cancer [10]. Subsequently, the principle commitment of this paper is to examine the adequacy of the recommended strategy in diagnosing the lymph infection issue

A few methodologies have been researched utilizing traditional and artificial intelligence procedures so as to assess the lymphography dataset. Karabulut et al. considered the impact of highlight choice methods with NaïveBayes, Multilayer Perceptron (MLP), and J48 choice tree classifiers with fifteen genuine datasets including lymph infection dataset [11]. The best precision was 84.46% accomplished utilizing Chi-square FS and MLP. Derrac et al. proposed a developmental calculation for data decrease upgraded by Unpleasant set-based element choice. The best precision recorded was 82.65% with 5 neighbors [12]. Incense [13] proposed a relative report between Naive Bayes, Tree Expanded Naive Bayes (TAN) and General Bayesian system (GBN) classifier, with K2 hunt and GBN with slope climbing seek in which they scored a precision of 82.16%, 81.07%, 77.46% and 75.06% individually. De Falco [14] proposed a differential advancement system to group eight databases from the medical space. The recommended strategy scored a precision of 85.14% contrasted with 80.18% utilizing Part classifier. Abellán and Masegosa structured Packing credal choice trees utilizing loose probabilities and vulnerability measures. The proposed choice tree show without pruning scored an exactness of 79.69% and 77.51% with pruning [15].

In this article, a CAD system dependent on Naïve Bayes Tree (NBTree) classifier is acquainted with improves the proficiency of the characterization exactness for lymph

sickness diagnosis. The distinction between this article and different articles that address a similar point is that a solid outfit classifier plot has been made by the use of NBTree, which yields more effective outcomes than any of different methods tried in this paper. We saw a promising improvement in arrangement execution of the calculation with resampling methodology.

The article starts with the recommended highlight choice methods and the NBTree classifier in Section 2. Segment 3 portrays the investigation steps and the included dataset and demonstrates the aftereffect of the analyses. The article finishes up with end and further research in Segment 4.

**2. Proposed Method**

The NBTree method that we use is a half and half of two classifiers, the ID3 choice tree and Naive Bayes. ID3 is fascinating in its portrayal of information, its way to deal with the administration of multifaceted nature, its heuristic for selecting candidate concepts, and its potential for taking care of loud data. It speaks to the idea of choice tree that take into account arrangement for an item by testing its incentive for specific properties. The algorithm of ID3 choice tree is as per the following:

```
function induce_tree (children_set, RiskFactors)
   begin
   if all entries in children_set are in the same class
    then return a leaf node labeled with that class
    else if RiskFactors is empty
       then return leaf node labeled with disjunction
       of all classes in children_set
       else begin
          select a property, P, and make it the root of the
          current tree;
          delete P from RiskFactors;
              for each value, V, of P,
               begin
                  create a branch of the tree labeled with V;
                  let partitionv be elements of children_set with
                  values V for property P;
                  call induce_tree (partition, RiskFactors), attach
                  result to branch V
                 end
        end
   end
```

The youngsters set in the calculation are the arrangement of the kid data while Hazard Variables is the Hazard factors that were utilized for grouping. ID3 applies the instigate tree work recursively to each segment. ID3 will begin by picking a hazard factor to be the base of the tree and keep developing branches and leaf. The leaf nodes of the three were included with Naive Bayes classifiers, rather than a node arranging a solitary class. Fig. 1 demonstrates the half and half method of choice tree and Naive Bayes classifiers on an example tree of youngster stoutness.

The Naive Bayes classifier depends on the Bayesian hypothesis and is especially appropriate for high measurement inputs. It is less complex than most methods yet despite everything it beats other refined characterization procedures and the illustration of NBTree can be given in Fig 1.

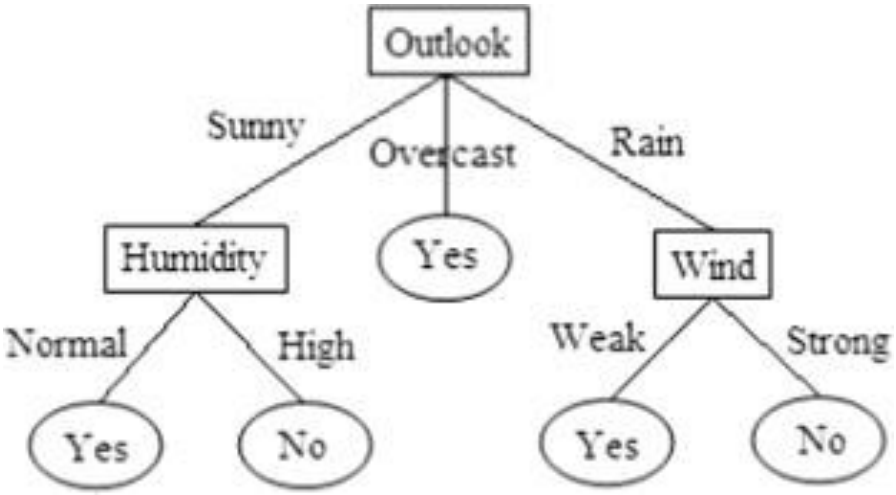


Fig. 1. Hybrid method of decision tree and NB

**3. Performance Validation**

The experiments are done in the WEKA tool, a familiar tool for of data mining algorithms written in Java. The dataset details are given in Table 1 and the output obtained from the WEKA tool is shown in Fig. 2.

Table 1 Dataset Description [16]

| Attribute number | Attribute description | Possible values of attributes | Assigned values | Mean | S.D. |
|---|---|---|---|---|---|
| 1 | Lymphatic | Normal = 1, arched = 2, deformed = 3, displaced = 4 | 1 - 4 | 2.74 | 0.82 |
| 2 | Block of afferent | No, Yes | 1 - 2 | 1.55 | 0.50 |
| 3 | Block of lymph c (superior and inferior flaps) | No, Yes | 1 - 2 | 1.17 | 0.38 |
| 4 | Block of lymph s (lazy incision) | No, Yes | 1 - 2 | 1.04 | 0.21 |
| 5 | By pass | No, Yes | 1 - 2 | 1.24 | 0.43 |
| 6 | Extravasates (force out of lymph) | No, Yes | 1 - 2 | 1.51 | 0.50 |
| 7 | Regeneration | No, Yes | 1 - 2 | 1.07 | 0.25 |
| 8 | Early uptake | No, Yes | 1 - 2 | 1.7 | 0.46 |
| 9 | Lymph nodes diminish | 0 - 3 | 0 - 3 | 1.06 | 0.31 |
| 10 | Lymph nodes enlarge | 1 - 4 | 1 - 4 | 2.47 | 0.84 |
| 11 | Changes in lymph | Bean = 1, oval = 2, round = 3 | 1 - 3 | 2.4 | 0.57 |
| 12 | Defect in node | No = 1, lacunar = 2, lacunar marginal = 3, lacunar central = 4 | 1 - 4 | 2.97 | 0.87 |
| 13 | Changes in node | No, lacunar, lacunar marginal, lacunar central | 1 - 4 | 2.8 | 0.76 |
| 14 | Changes in structure | No, grainy, drop-like, coarse, diluted, reticular, stripped, faint | 1 - 8 | 5.22 | 2.17 |
| 15 | Special forms | No, Chalices, vesicles | 1 - 3 | 2.33 | 0.77 |
| 16 | Dislocation | No, Yes | 1 - 2 | 1.67 | 0.48 |
| 17 | Exclusion of node | No, Yes | 1 - 2 | 1.8 | 0.41 |
| 18 | Number of nodes | 0 - 80 | 1 - 8 | 2.6 | 1.91 |
| 19 | Target Class | Normal = 1, metastases = 2, malign lymph = 3, fibrosis = 4 | | | |

Table 2 shows the performance of the NBTree interms of three measures namely precision, ROC and kappa.

**Table 2** Performance Evaluation of Different Classifiers

| Methods | Precision | ROC | Kappa |
|---|---|---|---|
| Proposed | 95.30 | 97.90 | 91.06 |
| Random Forest | 90.70 | 93.50 | 83.28 |
| k-NN | 92.70 | 75.40 | 87.13 |
| MLP | 90.70 | 91.40 | 83.28 |
| C4.5 | 88.80 | 78.50 | 79.50 |

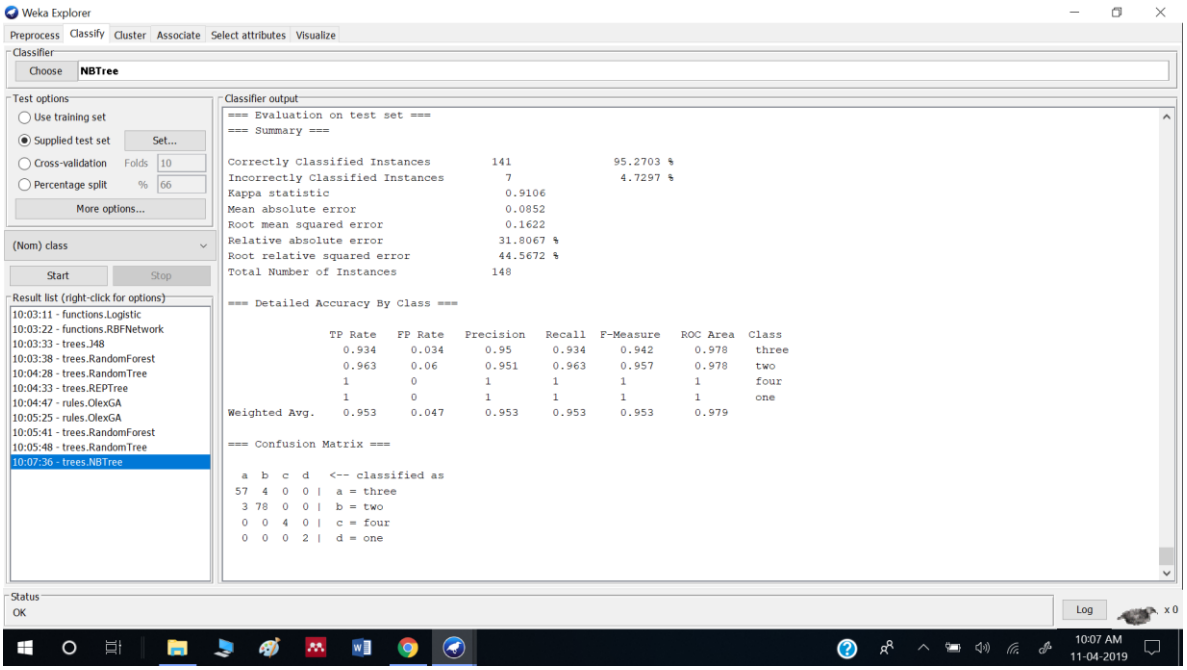


Fig. 2. Results obtained by proposed method in WEKA

From the table, it is clear that the maximum result is attained by the NBTree with the precision of 95.30. But, the C4.5 shows worst classification with the precision of 88.80. In addition, the MLP and RF exhibited identical results of 90.70 value of precision. These values looks better than the C4.5 method, but not than the NBTree and k-NN. Similarly, the k-NN shows better results with the maximum precision value of 92.70 which is better than all classifiers except NBTree. At the end, the presented NBTree classifier attains maximum results with the higher precision value of 95.30.

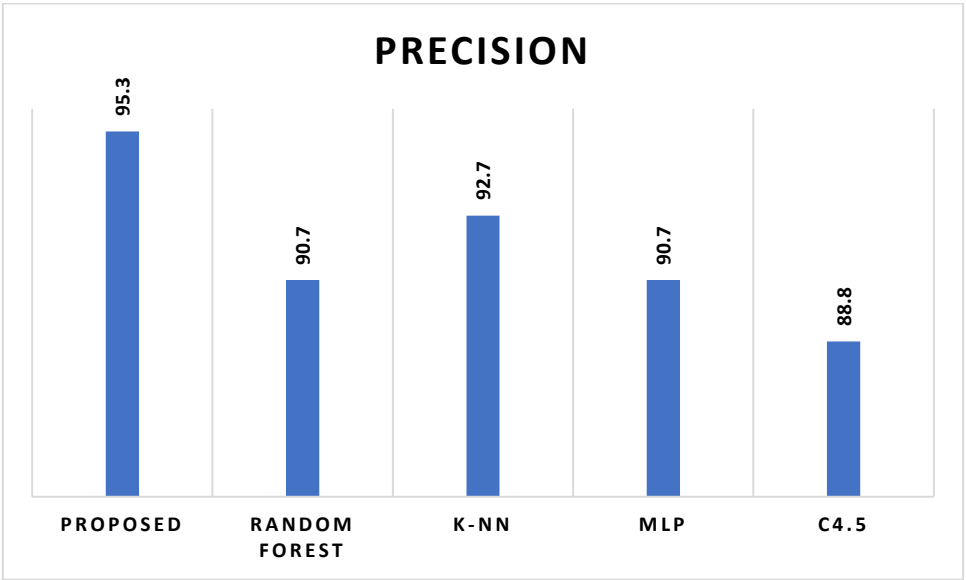


Fig. 2. Result analysis interms of precision

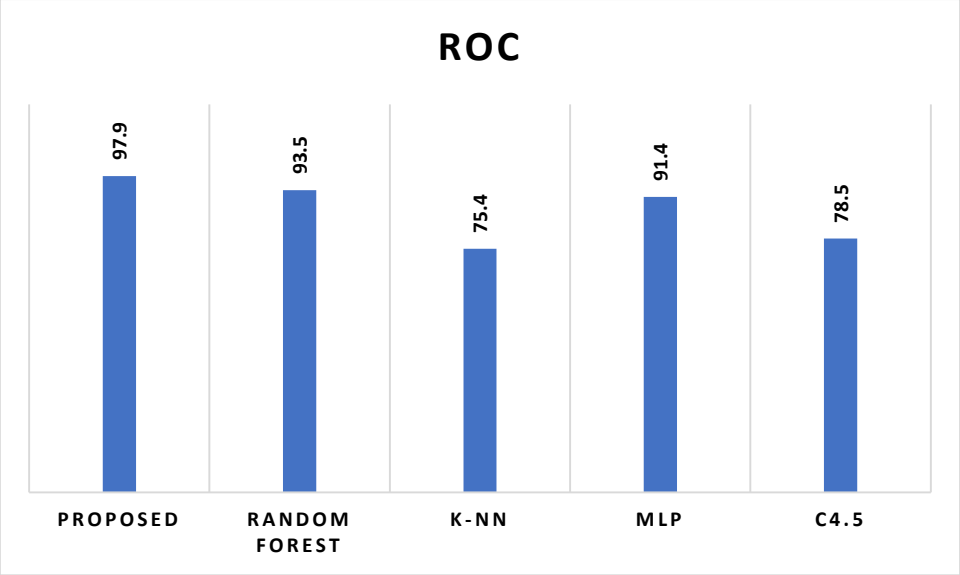


Fig. 3. Result analysis interms of ROC

From the Fig. 3, it is clear that the maximum result is attained by the NBTree with the ROC of 97.90. But, the C4.5 shows worst classification with the ROC of 78.50. In addition, the MLP and RF exhibited 91.40 and 93.50 values of ROC. These values looks better than the C4.5 method, but not than the NBTree and k-NN. Similarly, the k-NN shows poor results with the minimum ROC value of 75.40 which is better than all classifiers except NBTree. At the end, the presented NBTree classifier attains maximum results with the higher ROC value of 97.90.

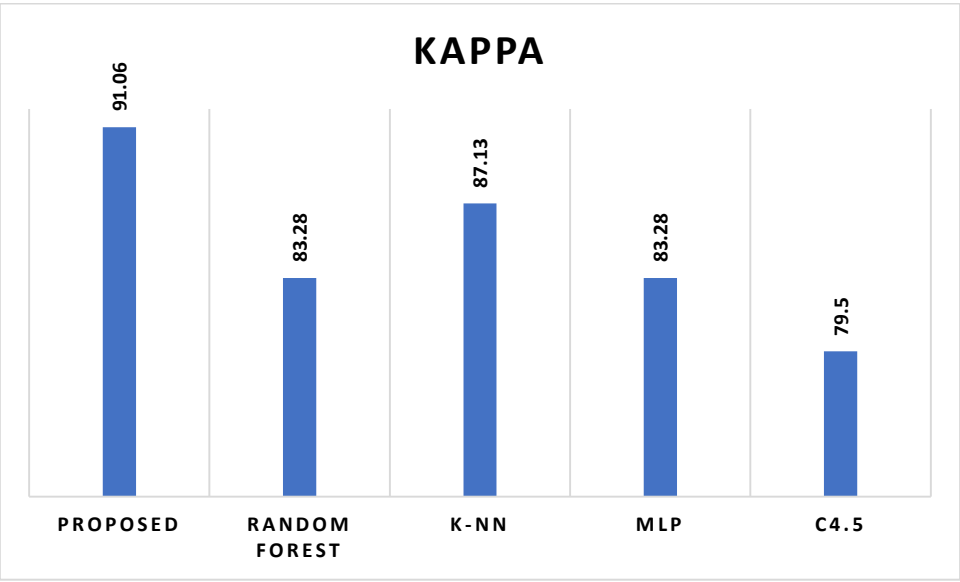


Fig. 4. Result analysis interms of kappa

From the Fig. 4, it is clear that the maximum result is attained by the NBTree with the kappa of 91.06. But, the C4.5 shows worst classification with the kappa of 79.50. In addition, the MLP and RF exhibited 91.40 and 93.50 values of KAPPA. These values looks better than the C4.5 method, but not than the NBTree and k-NN. Similarly, the k-NN shows poor results with the minimum KAPPA value of 75.40 which is better than all classifiers except NBTree. At the end, the presented NBTree classifier attains maximum results with the higher kappa value of 91.06

## 4. Conclusion

In this article, a CAD system dependent on Naïve Bayes Tree (NBTree) classifier is acquainted with improves the proficiency of the characterization exactness for lymph sickness diagnosis. The distinction between this article and different articles that address a similar point is that a solid outfit classifier plot has been made by the use of NBTree, which yields more effective outcomes than any of different methods tried in this paper. We saw a promising improvement in arrangement execution of the calculation with resampling methodology.

## References

[1] Ciosa, K.J. and Mooree, G.W. (2002) Uniqueness o Medical Data Mining. Artificial Intelligence in Medicine, **26**, l-24.

[2] Ceusters, W. (2000) Medical Natural Language Understanding as a Supporting Technology for Data Mining in Healthcare Medical Data Mining and Knowledge Discovery. Cios KJ Editor, Heidelberg: Springer, pp. 32-60,.

[3] Calle-Alonso, F., Pérez, C.J., Arias-Nicolás, J.P. and Martín, J. (2012) Computer-Aided Diagnosis System: A Bayesian Hybrid Classification Method. Computer Methods and Programs in Biomedicine, **112**, 104-113.

[4] Cselényi, Z. (2005) Mapping the Dimensionality Density and Topology of Data: The Growing Adaptive Neural Gas.Computer Methods and Programs in Biomedicine, **78**, 141-156. http://dx.doi.org/10.1016/j.cmpb.2005.02.001

[5] Huang, S.H., Wulsin, L.R., Li, H. and Guo, J. (2009) Dimensionality Reduction for Knowledge Discovery in Medical Claims Database: Application to Antidepressant Medication Utilization Study. Computer Methods and Programs in Biomedicine, **93**, 115-123. http://dx.doi.org/10.1016/j.cmpb.2008.08.002

[6] Luukka, P. (2011) Feature Selection Using Fuzzy Entropy Measures with Similarity Classifier. Expert Systems with Applications, **38**, 4600-4607. http://dx.doi.org/10.1016/j.eswa.2010.09.133

[7] Uthayakumar, J., Metawa, N., Shankar, K. and Lakshmanaprabu, S.K., 2018. Intelligent hybrid model for financial crisis prediction using machine learning techniques. *Information Systems and e-Business Management*, pp.1-29.

[8] Uthayakumar, J., Metawa, N., Shankar, K. and Lakshmanaprabu, S.K., 2018. Financial crisis prediction model using ant colony optimization. *International Journal of Information Management*.

[9] Guermazi, A., Brice, P., Hennequin, C. and Sarfati, E. (2003) Lymphography: An Old Technique Retains Its Usefulness. RadioGraphics, **23**, 1541-1558. http://dx.doi.org/10.1148/rg.236035704

[10] Cancer Research UK. http://www.cancerresearchuk.org

[11] Karabulut, E.M., Özel, S.A. and İbrikçi, T. (2012) A Comparative Study on the Effect of Feature Selection on Classification Accuracy. Procedia Technology, **1**, 323-327. http://dx.doi.org/10.1016/j.protcy.2012.02.068

[12] Derrac, J., Cornelis, C., García, S. and Herrera, F. (2012) Enhancing Evolutionary Instance Selection Algorithms by
Means of Fuzzy Rough Set Based Feature Selection. Information Sciences, **186**, 73-92.

[13] Madden, M.G. (2009) On the Classification Performance of TAN and General Bayesian Networks. Knowledge-Based Systems, **22**, 489-495. http://dx.doi.org/10.1016/j.knosys.2008.10.006

[14] De Falco, I. (2013) Differential Evolution for Automatic Rule Extraction from Medical Databases. Applied Soft Computing, **13**, 1265-1283. http://dx.doi.org/10.1016/j.asoc.2012.10.022

[15] Abellán, J. and Masegosa, A.R. (2012) Bagging Schemes on the Presence of Class Noise in Classification. Expert Systems with Applications, **39**, 6827-6837. http://dx.doi.org/10.1016/j.eswa.2012.01.013

[16] UCI (2016) Machine Learning Repository. http://archive.ics.uci.edu/ml/index.html