International Journal of Engineering & Technology, 7 (4.7) (2018) 297-302

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Hybrid Thyroid Stage Prediction Models Combining Classification, Clustering and Ensemble Systems

K.Pavya[1], Dr.B.Srinivasan[2]

[1] *Assistant professor, Department of Computer Science, Vellalar college for women*
[2] *Associate professor, Department of Computer Science, Gobi arts and science college*
**Corresponding author E-mail: pavyavcw@gmail.com*

Abstract

Early and correct detection of thyroid disease is very important for correct and timely treatment. The need to increase the accuracy of detecting and classifying thyroid disease poses a great challenge not only to the research community but also to healthcare industries. Usage of machine learning algorithms for thyroid disease classification is an area of research that is gaining popularity for the past few years. Automatic thyroid disease computer aided system for diagnosing the disease requires sophisticated and effective algorithms to perform classification in an accurate and time efficient manner. As a solution to this demand, hybrid models that combine clustering and classification algorithms along with ensemble technology are proposed. Four category of thyroid disease prediction system are proposed. They are Clustering + Classification models, Classification + Classification Models, Clustering + Clustering Models and Classification + Clustering Models. Two types of ensembles, namely, homogeneous and heterogeneous, are also considered and analyzed. Performance evaluation showed that the Classification + Classification model based on the combination of SVM and heterogeneous KNN + SVM classifier produce highest prediction accuracy.

***Keywords*:** *Combining Clustering and Classification Algorithms, Expectation-Maximization Clustering, Hybrid Prediction Models, K-Means Clustering, KNN Classifier, SVM classifier, Thyroid Disease Diagnosis.*

1. Introduction

Thyroid disease (TD) is a study of Endocrinology, which is one of the most common disease that is frequently misunderstood and misdiagnosed. Usage of Computer Aided Diagnosis (CAD) systems along with data mining techniques [13][14] is gaining popularity for automatic diagnosis of TD. However, the task is extremely challenging as it is very difficult and tedious to detect. As early detection of the disease is most important for finding a better treatment plan, automatic detection and classification of the present stage of the disease plays an important role.

An Automatic Thyroid Disease Computer Aided Diagnosis (ATD-CAD) system performs classification (normal, hyper and hypo) in two major steps. They are, feature extraction and classification using machine learning algorithms. The feature extraction step collects or extracts features regarding the disease from various medical imaging systems like ultrasound. The second step uses machine learning classifiers to diagnose the disease. This step uses the features extracted to train the classifier, which is then used to predict the stage of thyroid disease. The result of classification can be used to analyze the present stage of the disease. Our previous work [15] presented a series of classifiers that can be used in this step. Frequently, another step, called dimensionality reduction or feature selection is used to improve the performance of ATD-CAD. Our previous publications [7][9] have analyzed various feature selection algorithms.

This paper, to further enhance the performance of ATD-CAD systems, presents prediction models that can predict the stage of thyroid disease of a patient more efficiently. Prediction of diseases are frequently performed using two machine learning algorithms, clustering and classification. Both have been successfully deployed in various applications individually [2]. Both the algorithms have their own merits and demerits. For example, eventhough classification algorithms are much preferred to clustering algorithms during prediction, their performance degrades when presented with small number of reliable labeled data. Moreover, these algorithms do not consider inter-dependencies between the data. Alternatively, clustering algorithms do not have label (target class) information and can consider data-relationships and therefore can provide additional constraints (like if two data objects are clustered together, then it is more likely that they have the same label) that can increase prediction accuracy of unknown data. Thus, systematic combination of these two types of machine learning algorithms can provide more merits during prediction and classification and is analyzed in this paper.

In general, hybridization is defined as the art of combining the two or more algorithms to solve a single problem by dynamically switching between the selected algorithms [5]. *These models work with the aim of increasing the performance and provide better prediction results.* Most of the existing hybrid systems combine two or more algorithms belonging to the same domain. For example, hybrid models combining different clustering algorithms or different classification algorithms [11].

Another novel way of forming hybrid models is to combine clustering and classification algorithms in order to improve prediction performance. Examples belonging to this category include the models proposed by [16] and [1]. These hybrid models combine a single clustering algorithm with a single classification algorithm. Several attempts have been made to enhance the working of these models, with the aim of improving its classification accuracy. In continuation with this line, this paper attempts to enhance

SCCHM with the use of Ensemble System (ES). ES has been frequently used both in supervised and unsupervised domains [6]. This paper analyzes the effectiveness of modifying SCCHM to combine the concepts of ensemble technology.

The usage of different learning algorithms or different instantiations of the same learning algorithm is termed as an ensemble system. An ensemble system allows different and difficult needs of a problem to be handled by algorithms that are best suited to their needs of the application. They have the advantage of providing an extra degree of freedom in the classical bias/variance tradeoff, thus allowing solutions that would be difficult (if not impossible) to reach with only a single learning algorithm. Because of these advantages, ensemble systems have been applied to many difficult real-world problems, like, statistics, machine learning and pattern recognition. Several studies have also compared the performance of ensemble classifiers with single classification system and have concluded that ensemble systems produce better results [3][12]. This paper proposes ensemble based hybrid model that combine clustering and classification algorithms. Four types of ensemble hybrid models are proposed, which differ in the manner in which the clustering and classification algorithms are combined. The four hybrid models are Clustering-based Ensemble Classification (C-ECL) model, Clustering-based Ensemble Clustering (C-EC) model, Classification-Based Ensemble Classification (CL-ECL) model and Classification-Based Ensemble Clustering (CL-EC) model. The common goal of all the four proposed types of hybrid models is to improve the prediction performance during thyroid disease identification. For this purpose, two clustering algorithms, K-Means (KM) and Expectation-Maximization (EM) algorithms and two classification algorithms, Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) are used. The main contribution of this paper is to find the best combination of the learning algorithms that can provide maximum benefit during thyroid disease classification and prediction.

The rest of the paper is organized as follows. Section 2 presents the methodology used to design the four proposed ensemble hybrid models. Section 3 presents and discusses the experimental results that evaluate the performance of the proposed models. Section 4 concludes the work with future research directions.

2. Methodology

As mentioned earlier, four types of hybrid ensemble models that combine ensemble concepts with clustering and classification algorithms are proposed. The ensemble clustering model works on the principal that if two features are grouped into the same cluster by more than one clustering algorithm, then they are highly likely to be in the same class. Similarly, the ensemble classification model makes sure that the final prediction does not deviate much from the majority voting of the classifiers. The proposed hybrid models are formed using two clustering (KM and EM) and two classification (SVM and KNN) algorithms.

An ECL model can be designed as either heterogeneous or homogenous. Homogeneous models are considered as those systems having the same learning methodology but different feature vectors, while heterogeneous models are models using models using different methodology. In the homogeneous version of prediction models, a specific number of instances of the same base learning method were used to create the ensemble. The random subspace selection algorithm is used during the creation of ensembles. This method, creates a fixed number of training and testing sets using a constant x that indicates the number of data that is to be used to create the two sets. In this work, x is set to 80%, that is, 80% of features are used as training feature set (TFS), while the rest 20% of the features are used as Testing Feature Set (TeFs). The partitioning is performed in a way that both the sets consist of records representing the three selected types of thyroid stages. The majority voting scheme is the aggregation method used to combine the results from the various base learning methods. Details regarding the ensemble models are presented in Table 1 and the six different ensemble models designed are listed below:

(i) Ensemble Classification (ECL) Models
a. Homogeneous ECL model designed with KNN classifier (HoKNN)
b. Homogeneous ECL model designed with SVM classifier (HoSVM)
c. Heterogeneous ECL model designed with KNN and SVM classifiers (HeKS)

(ii) Ensemble Clustering (EC) Models
a. Homogeneous EC model designed with 10 instances of K-Means clustering algorithm (HoKM)
b. Homogeneous EC model designed with 10 instances of EM clustering algorithm (HoEM)
c. Heterogeneous EC model designed with KNN and EM clustering algorithm (HeKE)

Table 1:. details on the design of ensemble models

Factors	Details
No. of Base Learning Algorithms	Heterogeneous : 2; Homogeneous : 10
Base Classifier Used	SVM and KNN
Base Clustering Algorithm Used	KM and EM
Ensemble Creation Methods	Random Subspace Selection Technique
Partitioning Method Used	Hold-out method
Aggregation Method	Majority Voting Algorithm

All the four proposed hybrid ensemble models perform thyroid disease stage prediction (normal, hyper and hypo) after feature selection algorithm as proposed in our previous works [8][10]. The optimal feature set obtained from feature selection step is divided into training and testing sets. As mentioned earlier, the thyroid disease prediction is performed using the proposed hybrid ensemble models.

In C-ECL and C-EC, the ensemble learning algorithm (ECL / EC) uses the result of clustering algorithm to reduce the training set, or in other words, this step filters out unrepresentative features. The steps involved are presented in Figure 1 (Solid Lines). The other two types of hybrid ensemble models are classification-based clustering models (CL-ECL and CL-EC) (Figure 1, Dashed Lines), where the classifier (either SVM or KNN) is used first, then the results are used by the clustering algorithm for prediction. As clustering is an unsupervised method, it does not have the capacity to distinguish data accurately when compared to supervised model. Therefore, a classifier is trained first, whose output is taken as input to the clustering algorithm. This will improve the output of clustering based prediction.

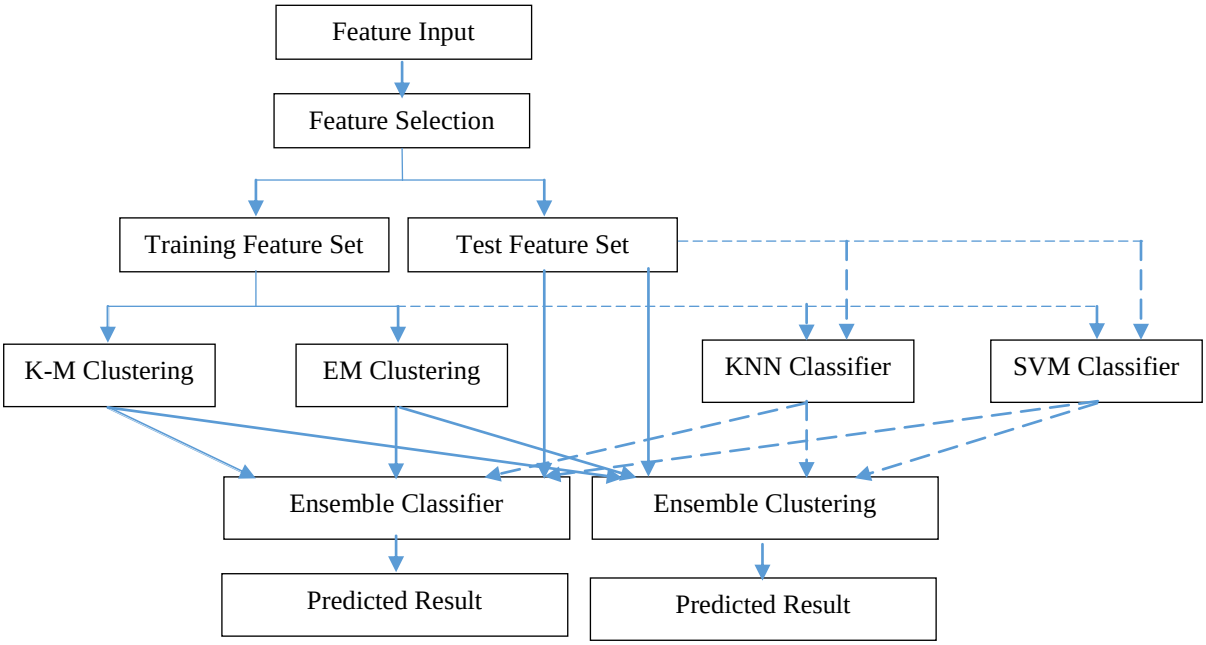


Figure 1 : Proposed Hybrid Models (Dashed Lines - C-Based Ensemble Models and Solid Lines CL-Based Ensemble Models)

2.1 Clustering Step of C-ECL and C-EC Models

Given a training feature set (TFS), the algorithm first uses a clustering algorithm (KM or EM) to extract cluster centers. The number of clusters is set to the number of classes to be classifies. As the aim of this work is to classify the stage of thyroid disease into three of its stages (that is, normal, hyper and hypo), the number of clusters is set to 3. In this step, each feature along with its nearest neighbour in the same cluster is also identified. This is done using the distance between a feature point (f_i) and all other feature points in the same cluster. Then, the shortest distance between two feature points (f_i and its nearest neighbour) is found. In the next step, a new distance measure using two values are calculated. The first is the sum of distance between all feature points and cluster centers (Dist1) and the second is the sum of distance between cluster centers and its nearest neighbour in the same cluster (Dist2). Using the sum of Dist1 and Dist2, the new distance value, Dist is calculated. This new distance acts as a new distance based feature set that represent each value in the TFS. Let this new training feature set be represented as TFS'. TFS' consists of features which are represented in a more efficient manner and also dimension_size(TFS') < dimension_size(TFS).

2.2 Classification Step of CL-ECL and CL-EC Models

This step initially uses the TFS to train the classifiers (SVM or KNN). All the data which are not correctly classified are considered as noise and are removed. The rest of the data, after removing the noisy data, is then considered as a new TFS having optimal features that have better discriminating power and can improve the performance of prediction of the learning algorithm as it includes only the positive results. Let this new feature set be denoted as TFS'. In the next step, TFS' is used to train the either another classifier (CL-ECL model) or clustered (CL-ECL). This trained classifier is then used to identify the stage of thyroid disease. The process is shown in Figure 2.

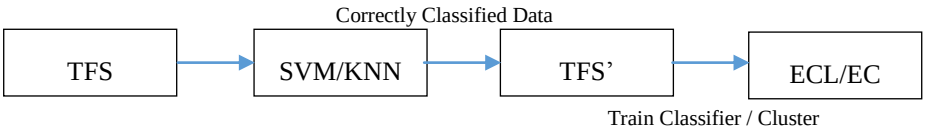


Figure 2 : Classification Process

2.3 Prediction of New Data using C-ECL and C-EC Models

In order to test (or predict the stage of) a new thyroid feature, the Test Feature Set (TeFS) is combined with the original with the original TFS. Next, the cluster heads and its nearest neighours are extracted to form the new set. In this step, the features in TeFS alone are considered. This step thus produces the new distance-based feature set (TeFS'). In C-ECL model, theTFS' and TeFS' are used to train and test the ensemble classifier for identifying the stage of thyroid disease of a patient. The C-EC model, after obtaining the clusters, during testing, estimates the cluster whose cluster head is close to the new thyroid data record. For this purpose, a distance measure (Euclidean distance) of the new data and to the cluster heads of the formed clusters is used and the cluster with minimum distance is selected and the new data is assigned to that group. The resultant cluster is then reported as the predicted class.

2.4 Prediction of New Data Using C-ECL and C-EC models

In CL-ECL model, the trained ensemble classifier is used to train the ensemble classifier, which is then used to predict the class of the new data. On the other hand, the reduced feature set TFS' is clustered using the ensemble clustering algorithm and when a new data is given as input, as with C-EC model, the closest cluster is identified and the new data is predicted to be in that stage of thyroid disease.

2.5 Proposed Hybrid Models

By varying the different type of clustering and classifiers used, different variants of the four proposed models were derived. Four homogeneous and two heterogeneous ensemble models were created which were then used to form six hybrid models under each type. Thus, a total of 24 hybrid models are proposed (Table 2). Here, Ho refers to homogeneous and He refers to heterogeneous type of ensemble.

Table 2. Proposed models

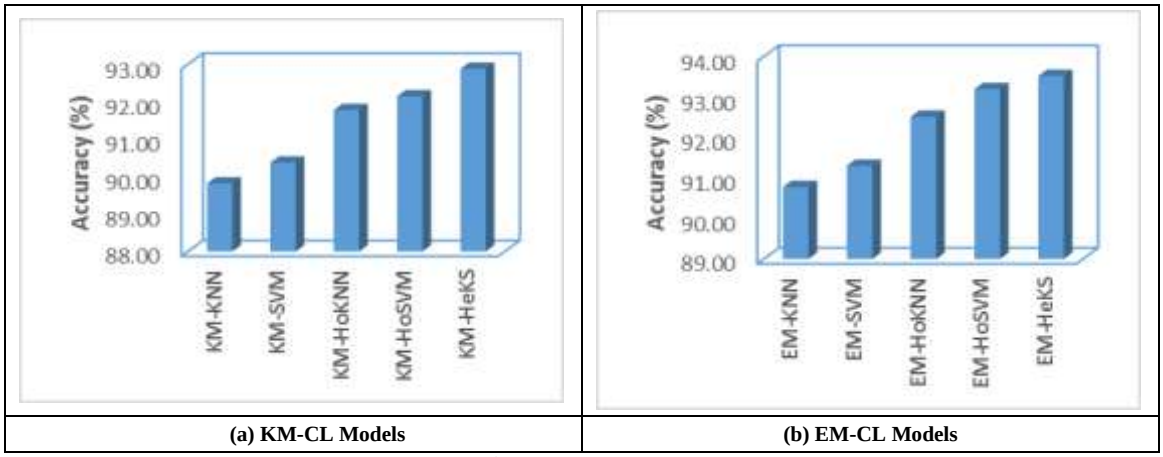
C-ECL		C-EC		CL-ECL		CL-EC	
KM-HoKNN	EM-HoKNN	KM-HoKM	EM-HoKM	SVM-HoKNN	KNN-HoKNN	SVM-HoKM	KNN-HoKM
KM-HoSVM	EM-HoSVM	KM-HoEM	EM-HoEM	SVM-HoSVM	KNN-HoSVM	SVM-Ho-EM	KNN-Ho-EM
KM-HeKS	EM-HeKS	KM-HeKE	EM-HeKE	SVM-HeKS	KNN-HeKS	SVM-HeKE	KNN-HeKE

3. Experimental Results

Performance evaluation of the proposed models was done using the thyroid disease database from UCI machine [4]. The dataset has details collected from 215 patients from the same hospital. The patients belonged to three groups of known classification and the class distribution is (i) Class 1: Healthy individuals (normal) - 150 individuals (ii) Class 2: Patients suffering from hyperthyroidism (hyper) - 35 individuals and (iii) Class 3: Patients suffering from hypothyroidism (hypo)- 30 individuals. In order to guarantee the valid results, the 10-fold cross validation method was used to evaluate the classification accuracy. The hybrid models were evaluated using the prediction accuracy performance metric.

3.1 Analysis of C-ECL Models

Figures 3a and 3b show the accuracy of the hybrid models (C-CL, C-HoECL and C-HeECL) that combine KM and EM clustering with different classification models respectively.



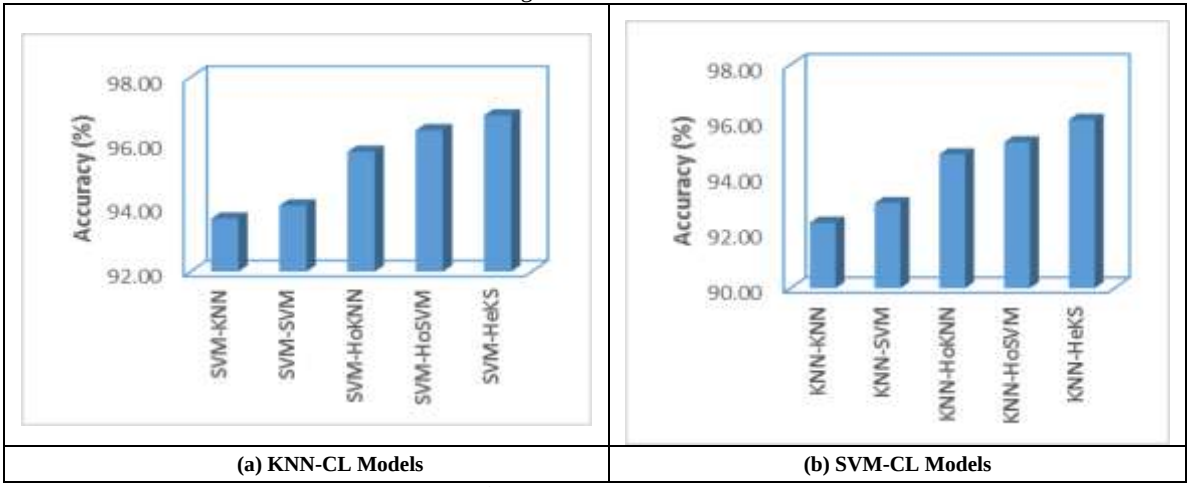
(a) KM-CL Models **(b) EM-CL Models**

Figure 3 : Accuracy Analysis of C-CL and C-HoECL and C-HeECL Hybrid Models

Comparative results reveals that the models that use EMclustering to train the classifiers yield better results when compared to models that use the result of KM. This indicates that the EM algorithm produces clusters with maximum quality and therefore, the ensemble classifiers trained by it provide more advantage during thyroid stage identification. The results further show that the hybrid models that combine clustering with heterogeneous ensemble classifiers (KM-HeKS – 92.88% and EM-HeKs – 93.52%) produces maximum classification performance. Comparison of homogeneous ensemble models show that the model combining EM clustering algorithm with homogeneous ensemble based on SVM classifier produces better classification results.

3.2 Analysis of CL-ECL Models

The results obtained by hybrid models that use the results of a classifier to train another classification model is shown in Figures 4a and 4b.



(a) KNN-CL Models **(b) SVM-CL Models**

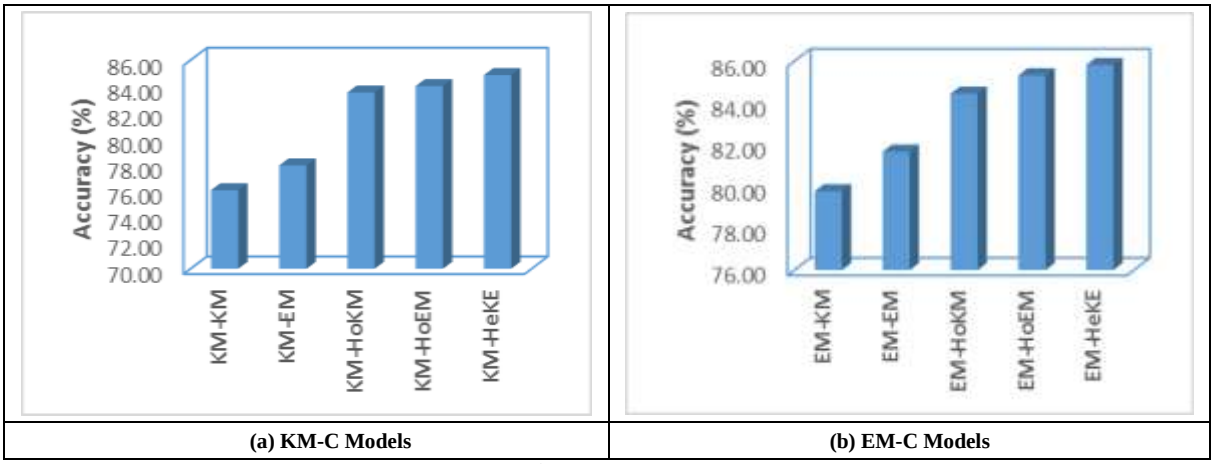
Figure 4 : Accuracy Analysis of CL-CL and CL-HoECL and CL-HeECL Hybrid Models

Based on the results presented in Table 2, mean age and the level of education in the control group are significantly deferent

The results pertaining to this category of hybrid models also show that the hybrid heterogeneous ensemble model classifier produces maximum efficiency, with respect to accuracy, when using the optimal training set produced by SVM/KNN classifier. On the other hand, comparing the various homogeneous variants of CL-CL models, the model that used SVM to train the prediction ensemble model classified thyroid disease in a more efficient manner than the variants that used KNN classifier. In summary, while all the proposed ensemble performed better than the conventional CL-CL models, the SVM-HeKS model performed the best (96.82%).

3.3 Analysis of C-EC Models

Figures 5a and 5b presents the prediction results of the hybrid models in C-C category.The trend obtained by the various C-C models is similar to that of C-CL models in the sense that all the models that used EM for obtaining the optimal training set performed better when compared to models that used KM. Among the homogeneous ensemble versions, the model that used EM for producing training vector and ensemble EM clustering algorithm for prediction performed the best in this category. However, the clear winner in this category is the EM-HeKE model, which produced the highest accuracy of 85.82%.

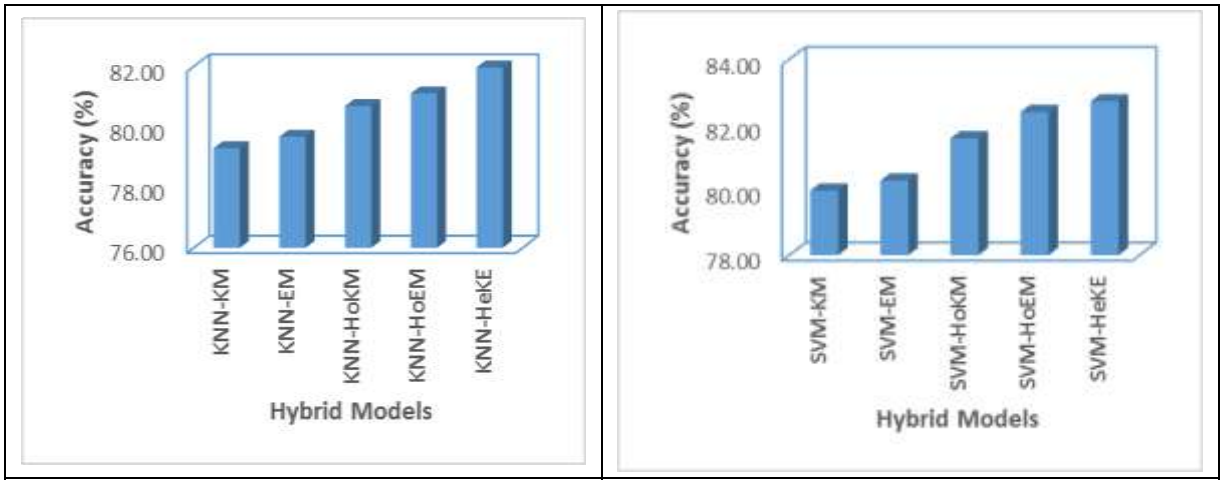


(a) KM-C Models **(b) EM-C Models**

Figure 5 : Accuracy Analysis of C-C and C-HoEC and C-HeEC Hybrid Models

3.4 Analysis of CL-EC Models

The analysis of the different CL-C models, with respect to accuracy performance metric, is shown in Figures 6a and 6b.



(a) KNN-C Models **(b) SVM-C Models**

Figure 6: Accuracy Analysis of CL-C and CL-HoEC and CL-HeEC Hybrid Models

As with CL-CL models, maximum benefit is given by CL-C models that used SVM classifier. Similar to the other three types of hybrid model, the performance of heterogeneous ensemble is better than homogeneous ensembles. Comparison of all the 10 proposed CL-C model reveals that the SVM-HeKE model performance is higher than all the other nine models.

3.5 Best Performing Model

Table 3compares the four best models from each category in terms of their capacity in predicting the stage of the thyroid disease.

Table 3:. Performance of the wining models in each category

Category	Model	Accuracy (%)
C-CL	EM-HeKS	93.52
CL-CL	SVM-HeKS	96.82
C-C	EM-HeKE	85.82
CL-EC	SVM-HeKE	82.72

Regarding this comparative results, the classification + heterogeneous classification model in the CL-CL category offers maximum accuracy benefits during classification (96.82%). This is followed by the C-CL category with 93.52% accuracy. The CL-EC model showed the minimum accuracy of 82.72%.

4. Conclusion

Using CAD during the diagnosis of thyroid disease is a maturing research field. The research focuses on improving the diagnosis system so that it can be helpful assistant tool to physicians during the detection of the disease and identification of the severity of the disease. This automatic system detects disease using two major steps, namely, feature extraction, selection and classification or prediction. This paper focus on algorithms that improve the prediction step. For this purpose, hybrid models are designed using a combination of clustering and classification algorithms using ensembling concepts. Four categories of hybrid models, namely, Clustering + Classification models, Classification + Classification Models, Clustering + Clustering Models and Classification + Clustering Models were designed. Two clustering algorithms (KM and EM) and two classification algorithms (KNN and SVM) were used during the implementation of these four categories. Six variants of hybrid models were created under each category by varying the selected clustering and classification algorithms. Further, both homogeneous and heterogeneous way of ensembling were studied and analyzed. Thus, a total of 24 hybrid models were implemented. Performance evaluation, using accuracy parameter, showed that the hybrid model that belonged to the Classification + Classification category produced maximum efficiency while using SVM classifier for training and KNN+SVM heterogeneous classifier for prediction. One general problem faced by ensemble systems are the high time complexity, solution to which is to be probed in future. Future research work is also planned in improving the working of the selected base line clustering and classification algorithms.

References

[1] Benvenuto, F., Piana, M. and Campi, C. and Massone, A.M. (2018) A Hybrid Supervised/Unsupervised Machine Learning Approach to Solar Flare Prediction, The Astrophysical Journal, Vol. 853, No. 1, Pp. 90-105.

[2] Chandra, B. (2009) Hybrid clustering algorithm, Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, IEEE Pp. 1345-1348.

[3] Dzelihodzic, A. and Donko, D. (2016) Comparison of Ensemble Classification Techniques and Single Classifiers Performance for Customer Credit Assessment, Modeling of Artificial Intelligence, Vol. 11, Issue 3, Pp. 140-150.

[4] http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease, Last Accessed During August, 2018.

[5] https://en.wikipedia.org/wiki/Hybrid_algorithm, Last Accessed During August, 2018.

[6] Hussein, S., Kandel, S., Bolan, C.W., Wallance, M.B. and Bagci, U. (2018) Supervised and Unsupervised Tumor Characterization in the Deep Learning Era, IEEE Transactions on Medical Imaging, Under Review, Pp. 1-11.

[7] Pavya, K. and Srinivasan, B. (2017a) Feature Selection algorithms to improve thyroid disease diagnosis, IEEE International Conference on Innovations in Green Energy and Healthcare Technologies, Pp. 1-5.

[8] Pavya, K. and Srinivasan, B. (2017b) Enhancing Filter Based Algorithms for Selecting Optimal Features from Thyroid Disease Dataset, International Journal of Advanced Research in Computer Science, Research Paper, Vol. 8, No. 9, Pp. 184-188.

[9] Pavya, K. and Srinivasan, B. (2018a) Review of Literature on Filter and Wrapper Methods for Feature Selection, International Journal of Engineering Sciences & Research Technology, Vol. 7, Issue 1, Pp. 137-143.

[10] Pavya, K. and Srinivasan, B. (2018b) Enhancing Wrapper Based Algorithms for Selecting Optimal Features from Thyroid Disease Dataset, Research Paper, International Journal of Computer Sciences and Engineering, Vol. 6, Issue 3, Pp. 7-13.

[11] Roy, S.S., Ahmed, M. and Akhand, M.A.H. (2018) Noisy image classification using hybrid deep learning methods, Journal of Information and Communication Technology, Vol. 17, No. 2, Pp. 233–269.

[12] Shen, H., Lin, Y., Tian, Q., Xu, K. and Jiao, J. (2018) A comparison of multiple classifier combinations using different voting-weights for remote sensing image classification, International Journal of Remote Sensing, Vol. 39, Issue 11, Pp. 3705-3722.

[13] Srinivasan, B. and Pavya, K. (2016a) A Study on Data Mining Prediction Techniques in Healthcare Sector, International Research Journal of Engineering and Technology, Vol. 3, Issue 3, Pp. 552-556.

[14] Srinivasan, B. and Pavya, K. (2016b) Diagnosis of Thyroid Disease Using Data Mining Techniques: A Study, International Research Journal of Engineering and Technology, Vol. 3, Issue: 11, Pp. 1191-1194.

[15] Srinivasan, B. and Pavya, K. (2016c) A Comparative Study on Classification Algorithms in Data Mining, International Journal of Innovative Science, Engineering & Technology, Vol. 3, Issue 3, Pp. 415-418.

[16] Tsai, C.F. and Chen, M.L. (2010) Credit rating by hybrid machine learning techniques, Elsevier Journal of Applied Soft Computing, Vol. 10, Pp. 374-380.