HMV: A medical decision support framework using multi-layer classifiers for disease prediction

Anuradha Bansal, Akanksha Dara, Maitri Shastri, Prateek Gupta, T Dinesh Ram Kumar, Vrinda Poddar

Birla Institute of Technology and Science, Pilani

Abstract— **Decision support is a crucial function for decision makers in many industries. Typically, Decision Support Systems (DSS) help decision-makers to gather and interpret information and build a foundation for decision-making. Medical Decision Support Systems (MDSS) play an increasingly important role in medical practice. By assisting doctors with making clinical decisions, DSS are expected to improve the quality of medical care. Conventional clinical decision support systems are based on individual classifiers or a simple combination of these classifiers which tend to show a moderate performance. In this paper, a multi-layer classifier ensemble framework is proposed based on the optimal combination of heterogeneous classifiers. The proposed model named "HMV" overcomes the limitations of conventional performance bottlenecks by utilizing an ensemble of seven heterogeneous classifiers. The framework is evaluated on a number of diseases' datasets obtained from public repositories. Effectiveness of the proposed ensemble is investigated by a comparison of results with several well-known classifiers as well as ensemble techniques.**

Index Terms— **Ensemble technique. AdaBoost, Majority Voting, HMV (Hierarchical Majority Voting), Disease Classification, Data Mining, Multi-layer**

I. INTRODUCTION

Data mining in medical domain, is a process of discovering hidden patterns and information from large medical datasets; analysing them and using them for disease prediction [1]. Significant amount of work has already been done on disease classification and prediction. However, there is no single methodology which is generic enough to show consistent results and good accuracies on various different diseases' datasets.

This paper aims at creating a model that is robust to the choice of dataset and to a single classification technique chosen by bringing in the concept of a multi-layer ensemble framework. The proposed research focuses on a novel combination of heterogeneous classifiers for disease classification and prediction, thus overcoming the limitations of individual classifiers. The novel combination of heterogeneous classifiers is presented which is Naïve Bayes, Linear Regression, Quadratic Discriminant Analysis, K-Nearest Neighbor, Support Vector Machine, Decision tree using Information Gain and Decision tree using the Gini Index. The multiple classifiers are used at multiple layers to further enhance disease prediction accuracy. An application has also been developed for disease

prediction. It is based on the proposed HMV ensemble framework. The proposed application can help both doctors and patients in terms of data management and disease prediction.

A. Background

A large number of predictive models can be developed from data mining techniques which enable classification and prediction tasks. The learning phase succeeds the feature extraction phase and can be categorized into supervised and unsupervised learning. Some examples of supervised learning include Artificial Neural Network (ANN), Support Vector Machine (SVM), and Decision Trees (DT). In unsupervised learning, there is no class label field in sample data. Examples include, K-mean clustering and Self-Organization Map (SOM). An ensemble approach performs better than the individual machine learning techniques by combining the results of individual classifiers [4,5]. There are multiple techniques that can be utilized for constructing the ensemble model and each result in different diagnosis accuracy (Bagging [6] and Boosting [7] are the most common methods).

B. Motivation

The major motivation behind choosing this paper was that not only is the the problem tackled by it extremely interesting, but it also has seemingly endless real-world applications. Moreover the methodology adopted and the features created are ingenious and intuitive at the same time. Significant amount of work has already been done on disease classification and prediction. However, there is no single methodology which shows highest performance for all datasets or diseases. Although one classifier shows good performance in a given dataset, another approach outperforms the others for some other dataset or disease. Since, the proposed model named “HMV” overcomes the limitations of conventional performance bottlenecks by utilizing an ensemble of seven heterogeneous classifiers, we were motivated to further explore the techniques presented in this paper and try to improvise the results presented.

C. Objective

Primarily, our project aims at analysing, implementing, remodelling and verifying the results published in [1]. The problem statement has been reproduced here for the sake of clarity:

“To create a medical decision support framework for disease prediction using a multi-layer classifier ensemble.”

Through this objective we aimed at gaining practical knowledge of different data mining techniques

and a general understanding of the problem solving approach for different data mining tasks. We further aimed at gaining some insights to different classification techniques, feature formation etc. The approach we used for disease classification using the HMV (Hierarchical Majority Voting) framework can be extended for being able to make predictions in several other domains as well.

II. Related work

Extensive amount of work has already been done on disease classification and prediction. However, most of the literature has focused on using a single classifier for a specific disease.

A neural networks ensemble method for prediction of heart diseases on Cleveland dataset has been proposed in [8]. Prashanth et al. [12] proposed automatic classification and prediction of Parkinson's disease. SVM and logistic regression are used for model construction. SVM classifier with RBF kernel produced high classification accuracy. Improvements can be made by incorporating ensemble classifier instead of a single classifier.

The paper [9] proposes a AptaCDSS-E system, clinical decision support system for cardiovascular disease level prediction, that overcomes conventional performance limitations by utilizing ensembles of different classifiers. Temurtas [10] introduced a neural network ensemble method for thyroid disease diagnosis in medical datasets. The proposed research focuses on using multilayer, probabilistic and learning vector quantization methods for implementing the neural networks. However, the framework is tested only for thyroid disease. In [11], a multi-layer perceptron-based decision support system is developed to support the diagnosis of heart diseases.

The literature review shows that multiple techniques that have been utilized for disease classification and prediction. However, there is no single methodology which shows highest performance for all datasets or diseases. The paper that we are going to implement focusses on multi-classifier and multi-layer ensemble framework for disease classification and prediction with high accuracy for all diseases and datasets.

HMV: A medical decision support framework using multi-layer classifiers for disease prediction

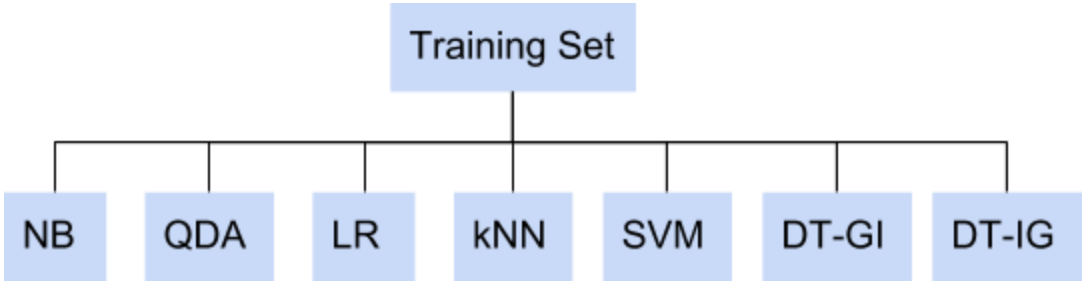
III. Proposed Technique(s) and Algorithm(s)

The technique proposed in the paper focuses on a novel combination of heterogeneous classifiers for disease classification and prediction: HMV (Hierarchical Majority Voting) ensemble framework.

The novel combination of heterogeneous classifiers to be used by HMV is:

1. Naïve Bayes (NB)
2. Linear Regression (LR)
3. Quadratic Discriminant Analysis (QDA)
4. K-Nearest Neighbor (KNN)
5. Support Vector Machine (SVM)
6. Decision tree using Information Gain (DT-IG)
7. Decision tree using the Gini Index (DT-GI)

The multiple classifiers are used at multiple layers to further enhance disease prediction accuracy. The ensemble classifier overcomes the limitations of individual classifiers as shown below.



The computational complexity of HMV ensemble framework is reduced by dividing it into three layer approach as follows:

Layer 1	NB, LR, QDA
Layer 2	Output of layer 1 using majority voting ensemble Two more classifiers at layer 2: SVM, KNN
Layer 3	Output of layer 2 is combined with DT-IG and DT-GI.

We also did several improvizations at each stage of the project.

HMV: A medical decision support framework using multi-layer classifiers for disease prediction

A. Pre-processing

Data acquisition and pre-processing module includes feature selection, missing value imputation, noise removal and outlier detection.

Data Acquisition: Created a generic function to load the datasets, replace the missing values in each of the datasets (denoted by "?", "0.0", "-9.0" etc.) with NaN, specify column headings for each attribute, to identify and return the set of attributes with missing values.

Data Normalization: it is done using MinMaxScaler. This estimator scales and translates each feature individually such that it is in the given range on the training set, i.e. between zero and one.
Computation of Missing Values: The kNN approach is used for missing data imputation. The proposed procedure is named as "kNNimpute". It is defined as, given a set of instances with incomplete pattern, the K closest cases with known attribute values are identified from the training cases for which the attribute values need to be determined. Once K-nearest neighbors are identified, the missing attribute values are then identified. Heterogeneous Euclidean-Overlap Metric (HEOM) is used for distance measure in order to determine the K-nearest neighbors and then to impute the missing value[13]. The process has been illustrated in the figure below (Source: [13]).

Removing Outliers: The outliers in medical data can exist due to several reasons such as abnormal condition of patient, equipment malfunction or recording error. The proposed HMV method uses Grubb's test for outlier detection and elimination from medical datasets. However, we have used the 1.5xIQR rule. This rule says that rule says that a data point is an outlier if it is more than 1.5.IQR above the third quartile or below the first quartile.

B. Feature Generation

Principal Component Analysis (PCA):
The principal component analysis technique assumes that most interesting and useful feature in the dataset is one which has the largest variance and spread [13]

F-score Feature Selection:
The HMV ensemble framework[13] utilizes F-score feature selection method in order to select most appropriate and relevant features from medical datasets. F-score method can distinguish between two classes having real values.

HMV: A medical decision support framework using multi-layer classifiers for disease prediction

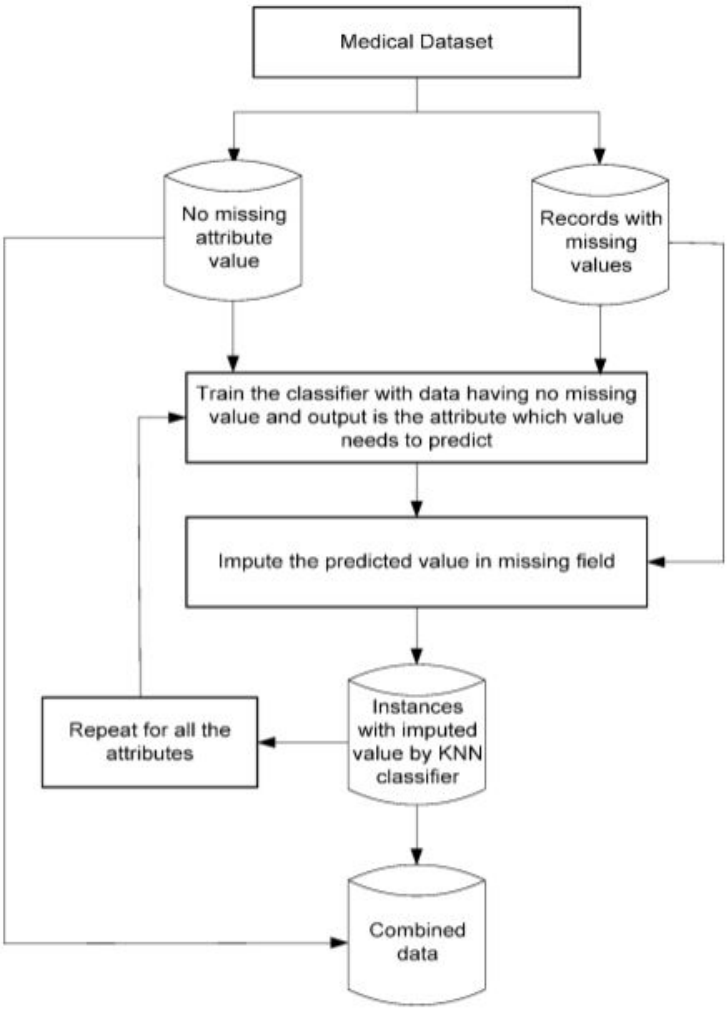


Fig. 1. Proposed missing data imputation method.

C. Classification (HMV Framework)

An ensemble classifier is considered to be of high quality if it able to give a high accuracy as well as prediction diversity. Weighted voting scheme considers the weight of each classifier. It combines the results of base classifiers and the ensemble model will output the class which has highest weight associated with it. The weights to each classifier can be assigned on the basis of classification accuracy. The highest weight will be assigned to the classifier which has highest accuracy. In case of unbalanced classes, the biasness of accuracy results can be generated due to biased dataset.

In order to avoid such situation, a multi-layer ensemble framework is proposed based on majority voting ensemble technique. The private judgment of each classifier is turned into collective decision by

aggregating their results. Multiple studies show that the strength of heterogeneous ensemble is related to the performance of the base classifiers and the lack of correlation between them (model diversity). The computational complexity of HMV ensemble framework is reduced by dividing it into three layer approach. The base classifiers at each layer are chosen in such a way that if any one classifier has some limitation, the other classifier performs well, consequently giving better performance.

IV. Datasets Used

The experimental evaluation of HMV ensemble framework is performed on two heart disease datasets, a breast cancer dataset, a diabetes dataset, two liver disease datasets, one hepatitis dataset and one Parkinson's disease dataset; which are as follows:

- Cleveland heart disease Dataset
- Statlog heart disease dataset
- Wisconsin Breast Cancer (WBC) dataset
- Pima Indian Diabetes Dataset (PIDD)
- BUPA Liver disease dataset
- Indian Liver Patient Dataset (ILPD)
- Hepatitis disease dataset
- Parkinson's disease dataset

Most of these datasets are publicly available and have been taken from the UCI data repository. The class labels of each dataset are replaced with 0 and 1 in order to maintain consistency where 0 represents absence of disease or healthy and 1 indicates the presence of disease or sick. Each dataset is divided into training set and test set. The HMV ensemble framework is applied on each test set. Ten-fold cross validation is applied and confusion matrices are obtained. The average prediction result of all confusion matrices is then calculated and analyzed.

V. Experiments and results

Application of the techniques proposed by the author along with our improvisations yielded some very intriguing results.

A. Methodology

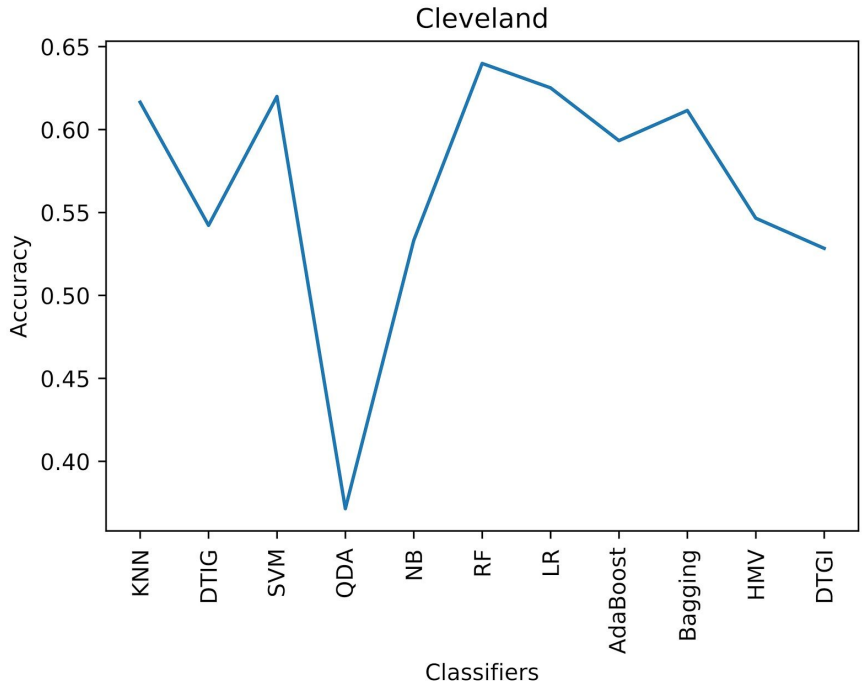
Various algorithms were applied on these feature sets to measure which feature set contributed towards the prediction to what extent. We measured the accuracy, specificity, sensitivity and F-measure scores for each of the 7 base classifiers, along with Adaboost, Random Forests (RF) and HMV for all the 8 datasets.

HMV: A medical decision support framework using multi-layer classifiers for disease prediction

B. Evaluation and Results

1. Accuracy table for Cleveland Heart disease Dataset

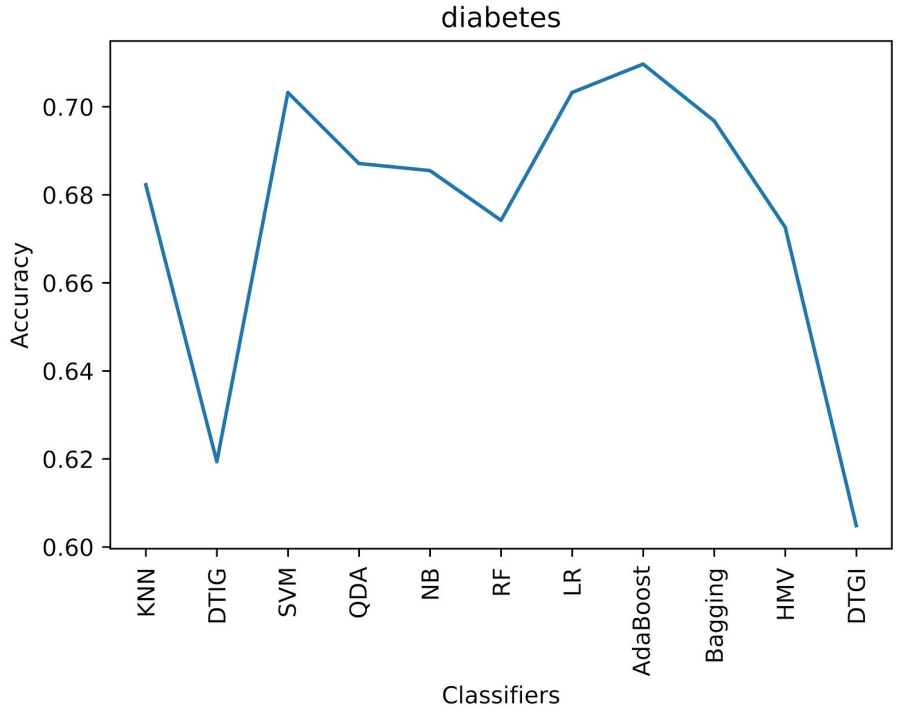
Classifier	Accuracy	Recall	Precision	F-Score
KNN	61.645021645021636	61.645021645021636	61.645021645021636	61.645021645021636
DTIG	54.220779220779214	54.220779220779214	54.220779220779214	54.220779220779214
SVM	61.991341991342	61.991341991342	61.991341991342	61.991341991342
QDA	37.14285714285714	37.14285714285714	37.14285714285714	37.14285714285714
NB	53.31168831168831	53.31168831168831	53.31168831168831	53.31168831168831
RF	63.98268398268399	58.8095238095238	58.419913419913414	61.53679653679654
LR	62.51082251082251	62.51082251082251	62.51082251082251	62.51082251082251
AdaBoost	59.32900432900432	59.32900432900432	59.32900432900432	59.32900432900432
Bagging	61.14718614718615	58.35497835497835	61.60173160173159	62.489177489177486
HMV	54.653679653679646	54.653679653679646	54.653679653679646	54.653679653679646
DTGI	52.83549783549783	52.83549783549783	52.83549783549783	52.83549783549783



HMV: A medical decision support framework using multi-layer classifiers for disease prediction

2. Accuracy table for Pima Indian Diabetes Dataset (PIDD)

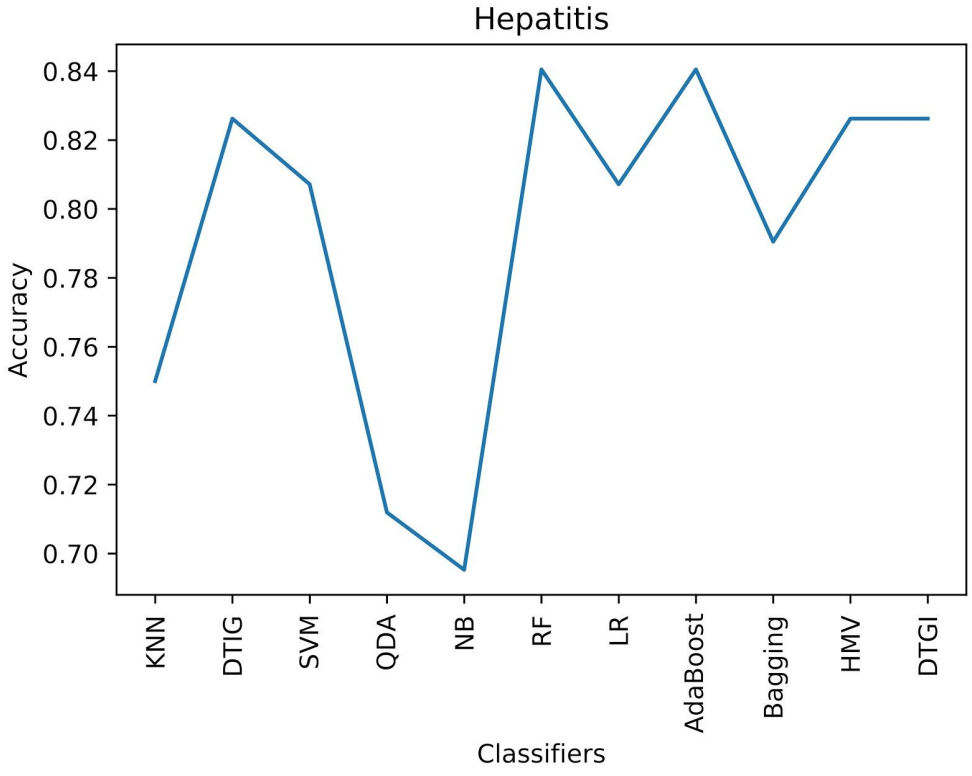
Classifier	Accuracy	Recall	Precision	F-Score
KNN	68.2258064516129	68.2258064516129	68.2258064516129	68.2258064516129
DTIG	61.935483870967744	61.935483870967744	61.935483870967744	61.935483870967744
SVM	70.3225806451613	70.3225806451613	70.3225806451613	70.3225806451613
QDA	68.70967741935485	68.70967741935485	68.70967741935485	68.70967741935485
NB	68.5483870967742	68.5483870967742	68.5483870967742	68.5483870967742
RF	67.41935483870968	65.80645161290323	67.0967741935484	68.06451612903226
LR	70.32258064516128	70.32258064516128	70.32258064516128	70.32258064516128
AdaBoost	70.96774193548387	70.96774193548387	70.96774193548387	70.96774193548387
Bagging	69.6774193548387	68.70967741935485	68.70967741935483	68.87096774193549
HMV	67.25806451612904	67.25806451612904	67.25806451612904	67.25806451612904
DTGI	60.483870967741936	60.483870967741936	60.483870967741936	60.483870967741936



HMV: A medical decision support framework using multi-layer classifiers for disease prediction

3. Accuracy table for Hepatitis Dataset

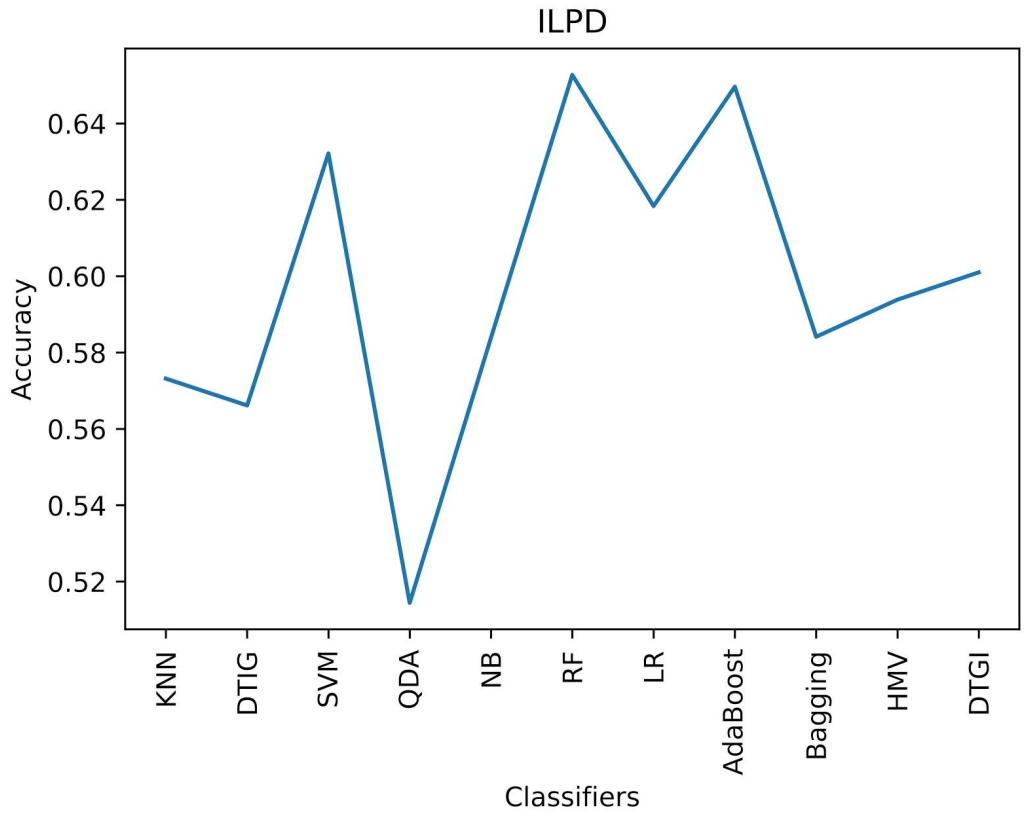
Classifier	Accuracy	Recall	Precision	F-Score
KNN	74.99999999999999	74.99999999999999	74.99999999999999	74.99999999999999
DTIG	82.61904761904762	82.61904761904762	82.61904761904762	82.61904761904762
SVM	80.71428571428572	80.71428571428572	80.71428571428572	80.71428571428572
QDA	71.19047619047618	71.19047619047618	71.19047619047618	71.19047619047618
NB	69.52380952380952	69.52380952380952	69.52380952380952	69.52380952380952
RF	84.04761904761905	82.38095238095238	84.04761904761905	82.38095238095238
LR	80.71428571428572	80.71428571428572	80.71428571428572	80.71428571428572
AdaBoost	84.04761904761905	82.61904761904762	82.61904761904762	84.04761904761905
Bagging	79.04761904761905	75.71428571428572	79.04761904761905	80.71428571428572
HMV	82.61904761904762	82.61904761904762	82.61904761904762	82.61904761904762
DTGI	82.61904761904762	82.61904761904762	82.61904761904762	82.61904761904762



HMV: A medical decision support framework using multi-layer classifiers for disease prediction

4. Accuracy table for ILPD

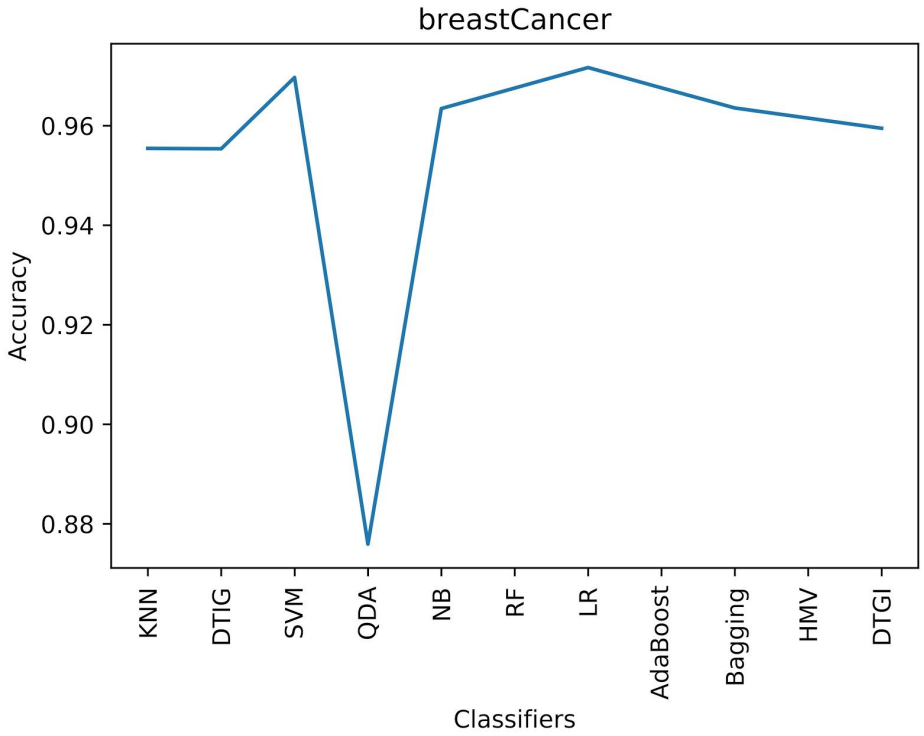
Classifier	Accuracy	Recall	Precision	F-Score
KNN	57.31527093596059	57.31527093596059	57.31527093596059	57.31527093596059
DTIG	56.61330049261084	56.61330049261084	56.61330049261084	56.61330049261084
SVM	63.21428571428572	63.21428571428572	63.21428571428572	63.21428571428572
QDA	51.44088669950738	51.44088669950738	51.44088669950738	51.44088669950738
NB	58.38669950738916	58.38669950738916	58.38669950738916	58.38669950738916
RF	65.27093596059113	61.1576354679803	62.51231527093596	65.62807881773399
LR	61.83497536945814	61.83497536945814	61.83497536945814	61.83497536945814
AdaBoost	64.96305418719213	64.96305418719213	64.96305418719213	64.96305418719213
Bagging	58.41133004926109	60.14778325123154	64.3103448275862	59.38423645320196
HMV	59.38423645320198	59.38423645320198	59.38423645320198	59.38423645320198
DTGI	60.09852216748768	60.09852216748768	60.09852216748768	60.09852216748768



HMV: A medical decision support framework using multi-layer classifiers for disease prediction

5. Accuracy table for Breast Cancer Dataset

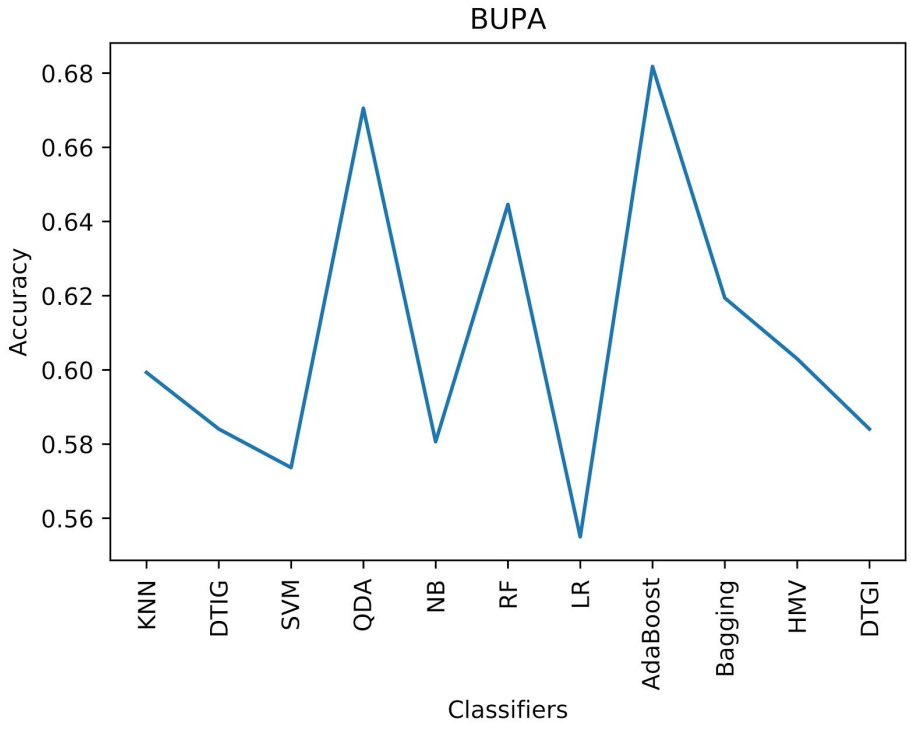
Classifier	Accuracy	Recall	Precision	F-Score
KNN	95.54285714285713	95.54285714285713	95.54285714285713	95.54285714285713
DTIG	95.53469387755102	95.53469387755102	95.53469387755102	95.53469387755102
SVM	96.9673469387755	96.9673469387755	96.9673469387755	96.9673469387755
QDA	87.59183673469387	87.59183673469387	87.59183673469387	87.59183673469387
NB	96.34285714285716	96.34285714285716	96.34285714285716	96.34285714285716
RF	96.75510204081633	96.55510204081634	96.75918367346938	96.15102040816326
LR	97.16734693877551	97.16734693877551	97.16734693877551	97.16734693877551
AdaBoost	96.7591836734694	96.7591836734694	96.7591836734694	96.7591836734694
Bagging	96.35510204081633	97.17142857142858	96.35102040816325	96.75102040816327
HMV	96.15102040816325	96.15102040816325	96.15102040816325	96.15102040816325
DTGI	95.94693877551019	95.94693877551019	95.94693877551019	95.94693877551019



HMV: A medical decision support framework using multi-layer classifiers for disease prediction

6. Accuracy table for BUPA dataset

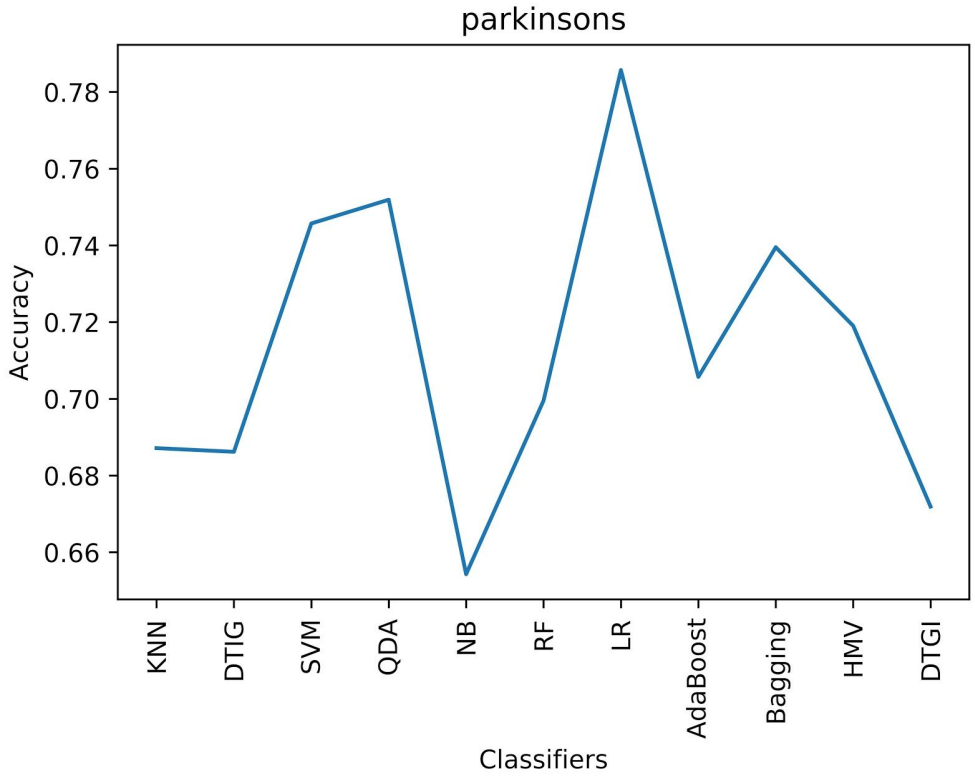
Classifier	Accuracy	Recall	Precision	F-Score
KNN	59.92877492877493	59.92877492877493	59.92877492877493	59.92877492877493
DTIG	58.404558404558415	58.404558404558415	58.404558404558415	58.404558404558415
SVM	57.36467236467235	57.36467236467235	57.36467236467235	57.36467236467235
QDA	67.05128205128204	67.05128205128204	67.05128205128204	67.05128205128204
NB	58.06267806267806	58.06267806267806	58.06267806267806	58.06267806267806
RF	64.45868945868946	67.76353276353277	66.62393162393163	64.01709401709402
LR	55.498575498575484	55.498575498575484	55.498575498575484	55.498575498575484
AdaBoost	68.17663817663818	68.17663817663818	68.17663817663818	68.17663817663818
Bagging	61.93732193732193	60.74074074074074	61.0968660968661	56.2962962962963
HMV	60.299145299145295	60.299145299145295	60.299145299145295	60.299145299145295
DTGI	58.404558404558394	58.404558404558394	58.404558404558394	58.404558404558394



HMV: A medical decision support framework using multi-layer classifiers for disease prediction

7. Accuracy table for Parkinson's Disease Dataset

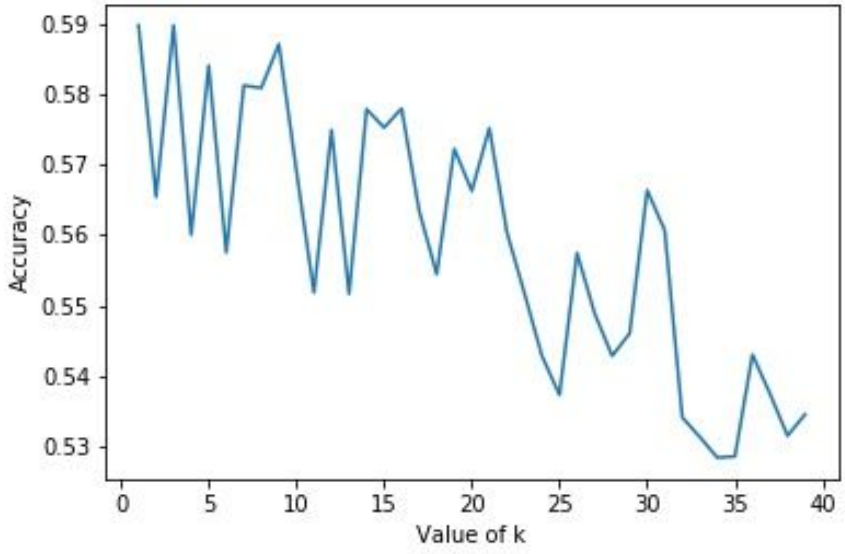
Classifier	Accuracy	Recall	Precision	F-Score
KNN	68.71428571428571	68.71428571428571	68.71428571428571	68.71428571428571
DTIG	68.61904761904762	68.61904761904762	68.61904761904762	68.61904761904762
SVM	74.57142857142858	74.57142857142858	74.57142857142858	74.57142857142858
QDA	75.19047619047619	75.19047619047619	75.19047619047619	75.19047619047619
NB	65.42857142857143	65.42857142857143	65.42857142857143	65.42857142857143
RF	69.95238095238095	68.57142857142857	68.61904761904762	67.28571428571429
LR	78.57142857142858	78.57142857142858	78.57142857142858	78.57142857142858
AdaBoost	70.57142857142857	70.57142857142857	70.57142857142857	70.57142857142857
Bagging	73.95238095238096	72.57142857142857	75.95238095238095	75.85714285714286
HMV	71.90476190476191	71.90476190476191	71.90476190476191	71.90476190476191
DTGI	67.19047619047619	67.19047619047619	67.19047619047619	67.19047619047619



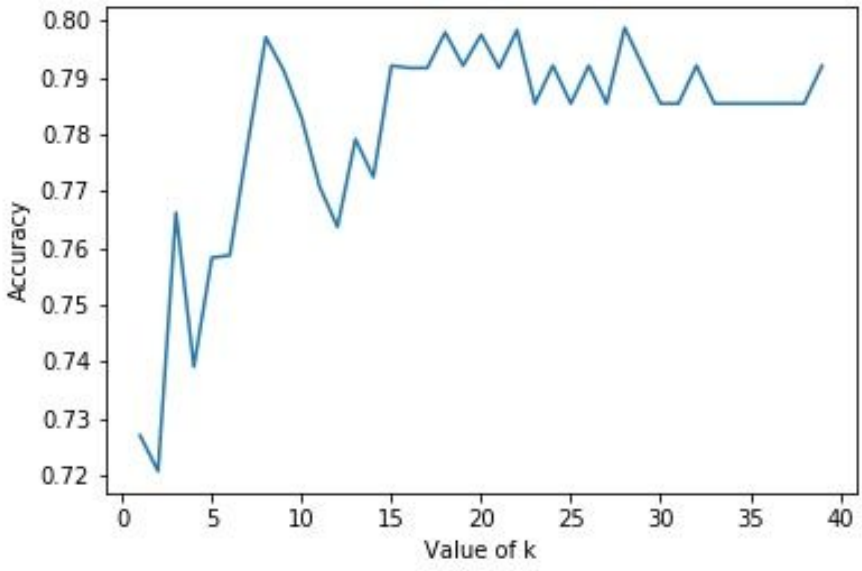
HMV: A medical decision support framework using multi-layer classifiers for disease prediction

GRAPHS FOR SELECTION OF K_KNN VALUES

BUPA dataset (k=3)

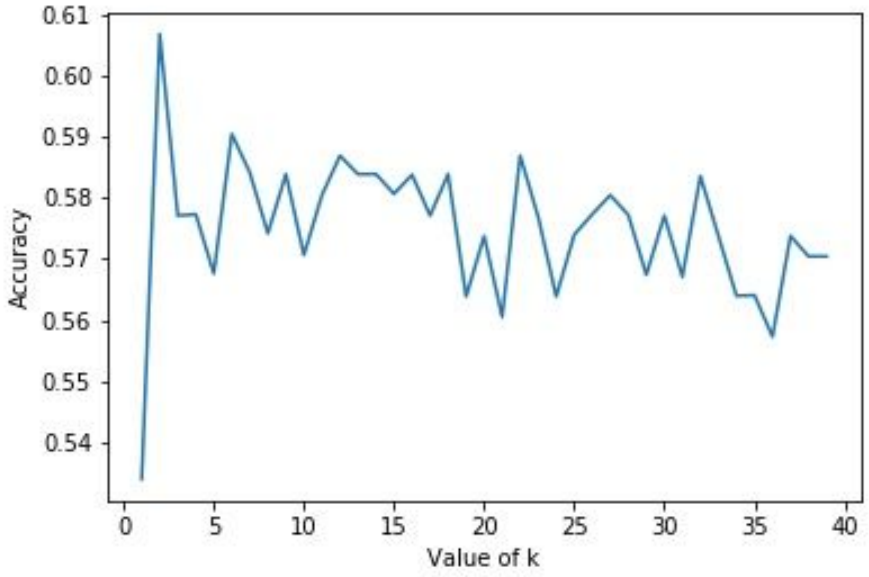


Hepatitis Dataset (k=10)

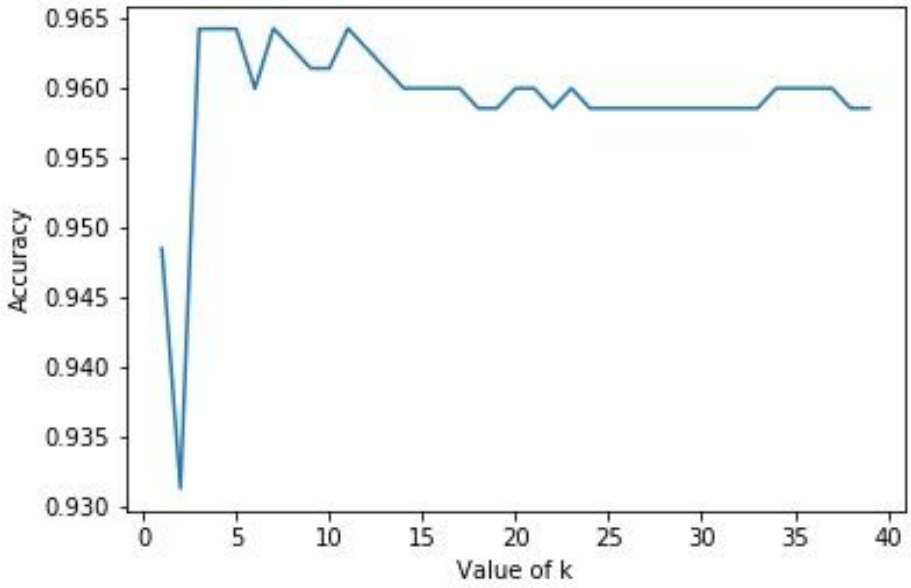


Cleveland Dataset (k=2)

HMV: A medical decision support framework using multi-layer classifiers for disease prediction

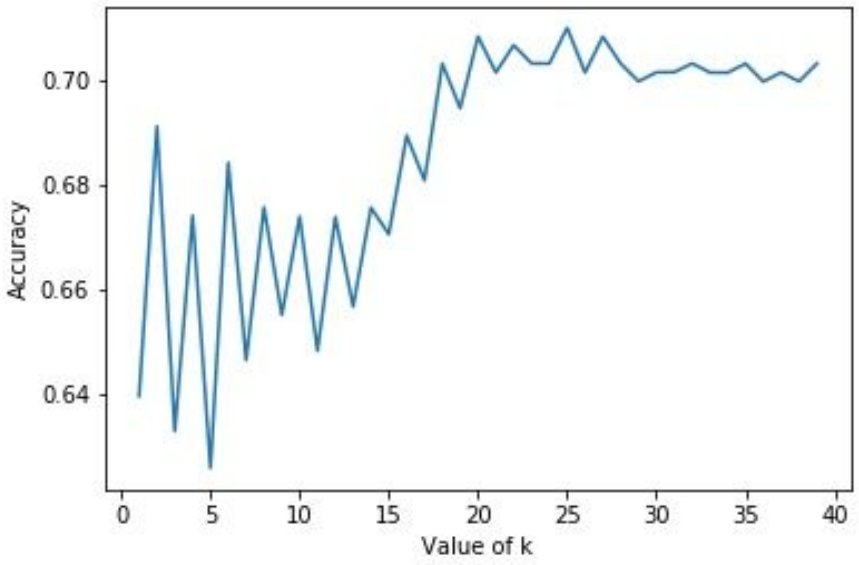


Breast Cancer Dataset (k=7)

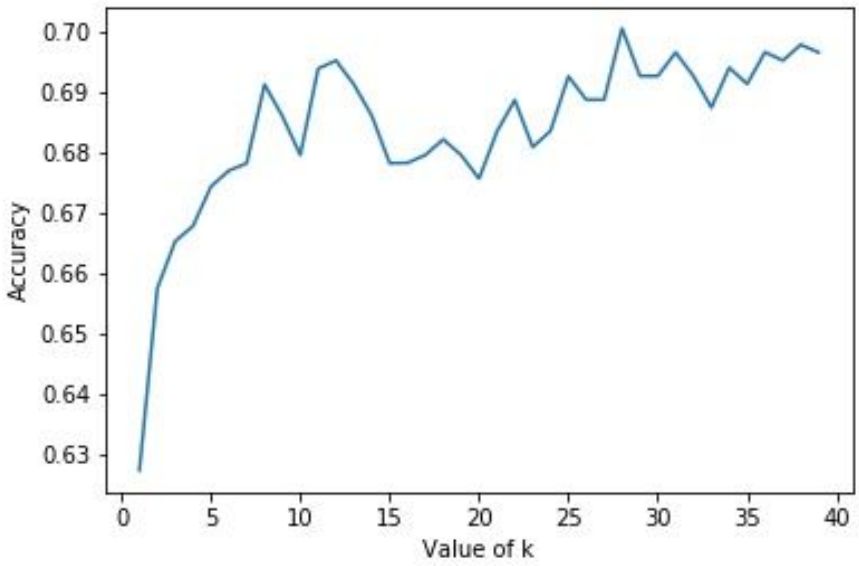


ILPD Dataset (k=25)

HMV: A medical decision support framework using multi-layer classifiers for disease prediction

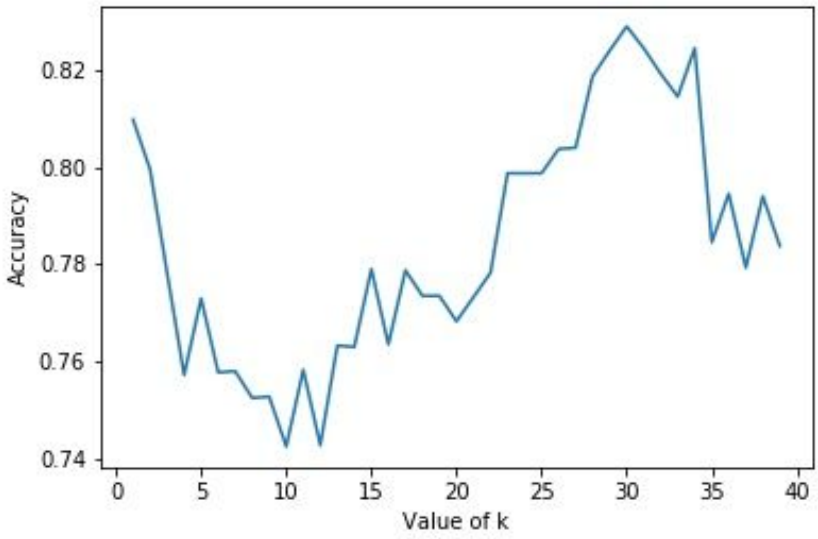


Diabetes Dataset (k=28)



Parkinson's Disease Dataset (k=30)

HMV: A medical decision support framework using multi-layer classifiers for disease prediction



VI. Conclusion and Future Work

Accuracy plays a vital role in the medical field as it concerns with the life of an individual. Data mining in the medical domain works on the past experiences and analyzes them to identify the general trends and probable solutions to the present situations. This research paper presents an ensemble framework using hierarchical majority voting and multi-layer classification for disease classification and prediction using data mining techniques. The proposed model overcomes the limitations of conventional performance by utilizing an ensemble of seven heterogeneous classifiers: Naïve Bayes (NB), Linear Regression (LR), Quadratic Discriminant Analysis (QDA), K Nearest Neighbor (kNN), Support Vector Machine (SVM), Decision tree using Information Gain (DT-IG) and Decision tree using Gini Index (DT-GI).

The proposed framework is based on three modules. The first module is data acquisition and preprocessing which obtains data from different data repositories and preprocess them. Each classifier's training is then performed on the training set in second module and then they are used to predict unknown class labels for test set instances. The prediction and evaluation is the third module of the proposed ensemble framework which is comprised of three classification layers. The evaluation of HMV framework is performed on two different heart disease datasets, two breast cancer datasets, two diabetes datasets, two liver disease datasets, one Parkinson's disease dataset and one hepatitis dataset

obtained from public repositories. The analysis of results indicates that proposed HMV ensemble framework has achieved highest accuracy of disease classification and prediction for all medical datasets. Moreover, a real-time implementation of proposed ensemble framework is also performed on blood CP dataset obtained from PIMS hospital in order to determine healthy and diseased patients. The analysis of results again shows high accuracy of disease prediction for real -time patients' data and also it can help practitioners and patients for disease prediction based on the disease symptoms.

Future Work

Currently, the proposed ensemble model predicts healthy and sick individuals based on their vital signs. It predicts either class 0 or class 1, representing either absence or presence of a disease. However, the proposed system can be extended to predict the levels and types of particular disease such as for heart disease, it can predict different levels of disease like early, acute, etc. and prediction of type-1 diabetes or prediction of type-2 diabetes in addition of predicting diabetes.

VII. References

[1] C. Fernández-Llatas, J.M. García-Gómez (Eds.), Data Mining in Clinical Medicine, Humana Press, 2015.

[2] J.H. Chen, T. Podchiyska, R.B. Altman, OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records, J. Am. Med. Inform. Assoc. (2015), Ocv091.

[3] S. Dua, X. Du, Data Mining and Machine Learning in Cyber Security, CRC Press, 2011.

[4] A. Ahmad, G. Brown, Random ordinality ensembles: ensemble methods for multi-valued categorical data, Inf. Sci. 296 (2015) 75–94.

[5] B. Sluban, N. Lavracˇ, Relating ensemble diversity and performance: a study in class noise detection, Neurocomputing 160 (2015) 120–131.

[6] F. Moretti, S. Pizzuti, S. Panzieri, M. Annunziato, Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling, Neurocomputing (2015).

[7] M.J. Kim, D.K. Kang, H.B. Kim, Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction, Expert Syst. Appl. 42 (3) (2015) 1074–1082.

[8] Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert systems with applications 36.4 (2009): 7675-7680.

[9] Eom, Jae-Hong, Sung-Chun Kim, and Byoung-Tak Zhang. "AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction." Expert Systems with Applications 34.4 (2008): 2465-2479.

HMV: A medical decision support framework using multi-layer classifiers for disease prediction

[10] F. Temurtas, A comparative study on thyroid disease diagnosis using neural networks, Expert Syst. Appl. 36 (1) (2009) 944–949.

[11] Yan, Hongmei, et al. "A multilayer perceptron-based medical decision support system for heart disease diagnosis." Expert Systems with Applications 30.2 (2006): 272-281.

[12] R. Prashanth, S.D. Roy, P.K. Mandal, S. Ghosh, Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging, Expert Syst. Appl. 41 (7) (2014) 3333–3342.

[13] Bashir, Saba, et al. "HMV: a medical decision support framework using multi-layer classifiers for disease prediction." Journal of Computational Science 13 (2016): 10-25.