Feature reduction improves classification accuracy in healthcare

By: Maha Asiri, Hamid Nemati, and Fereidoon Sadri

Asiri, Maha; Nemati, Hamid; Sadri, Fereidoon. 2018. Feature reduction improves classification accuracy in healthcare. IDEAS 2018: Proceedings of the 22nd International Database Engineering & Applications Symposium, June 2018. Pages 193–198.
https://doi.org/10.1145/3216122.3216165

Abstract:

Our work focuses on inductive transfer learning, a setting in which one assumes that both source and target tasks share the same features and label spaces. We demonstrate that transfer learning can be successfully used for feature reduction and hence for more efficient classification performance. Further, our experiments show that this approach increases the precision of the classification task as well.

Keywords: classification | transfer learning | feature reduction | classification accuracy

Article:

1 INTRODUCTION

The fields of data mining and machine learning have been widely and successfully used in many applications where patterns can be extracted from past information (training data) to predict future outcomes [13]. Machine learning has its advantages in all walks of life, with applications ranging from autonomous cars [3], cancer diagnosis [4, 9], heart disease prediction [10], and stock value prediction [12], to mention a few.

Usually, data is described by a set of features. We call a features unnecessary if it is either irrelevant to the current goal or holds redundant information given other features. Many machine learning algorithms tend to get overwhelmed when unnecessary features abound. They usually need more samples in the presence of irrelevant features. For example, the number of training samples needed for the basic nearest-neighbor classification algorithm to reach a given accuracy grows exponentially with the number of irrelevant features [6]. However, the success of supervised learning techniques depends on the presence of sufficiently large sets of training data. Ideally, these training sets are sampled from the same generating distribution that is expected to be present in production. Obtaining useful training sets is most often an arduous and expensive process. Transfer learning techniques [2, 5, 7, 8, 14] allow us to reuse knowledge (such as models or examples) gained from some learning task (called the source) and apply it to a related task for which enough training sets are not yet available (called the target). Effective transfer

learning techniques are much in need due to the growing demand for machine learning solutions for an ever-increasing number of computer applications and the tremendous growth in communicated information.

Consider, for example, the problem of automatic heart disease prediction using a health-tracking mobile app. The proportion of heart disease records is normally quite small in the context of any single mobile app user. Thus, the corresponding learning problem can be viewed as a classification problem with a small target class. Consequently, the acquisition of a sufficiently large labeled training set may take considerable time. If we already possess an annotated database of heart disease patient records from other users or hospitals, we could, hypothetically, use it for the current challenge (target) and benefit from transfer learning approach to utilize whatever source information is available, guided by the spares information already acquired for the new target.

Our work focuses on inductive transfer learning, a setting in which one assumes that both source and target tasks share the same features and label spaces. We demonstrate that transfer learning can be successfully used for feature reduction and hence for more efficient classification performance. Further, our experiments show that this approach increases the precision of the classification task as well.

The rest of this paper is organized as follows: We present our transfer learning methodology in Section 2. Then, In Section 3 we present our experimental results and comparisons, demonstrating the advantages of our inductive transfer learning approach. Our final observations and open questions for future research are provided with the concluding remarks in Section 4.

2 METHODOLOGY

2.1 Transfer learning Architecture

We introduce a simple algorithm for feature transfer learning. First, we identify a set of important features based on the source dataset, which are used to filter the dataset to obtain a target dataset that only contains these important features. Then we perform a second round of model construction and prediction on the reduced target dataset. Interestingly, this approach results in improved prediction accuracy.

To implement the above architecture, we followed the following steps:

- Pre-processing the dataset to replace string-valued features with numerical labels.
- Split data randomly into 80% for training and 20% for evaluation.
- Train 20 different models of Decision Tree and Random Forest based on a range of max depth parameter.
- Identify the top 10 important features for each run, and choose the top 10 features with highest frequency count.
- Filter the dataset to obtain the target dataset.
- Repeat the split, train, evaluate cycle.
- Obtain the best model for each algorithm.

We used Python scikit-learn [1] for our implementation.

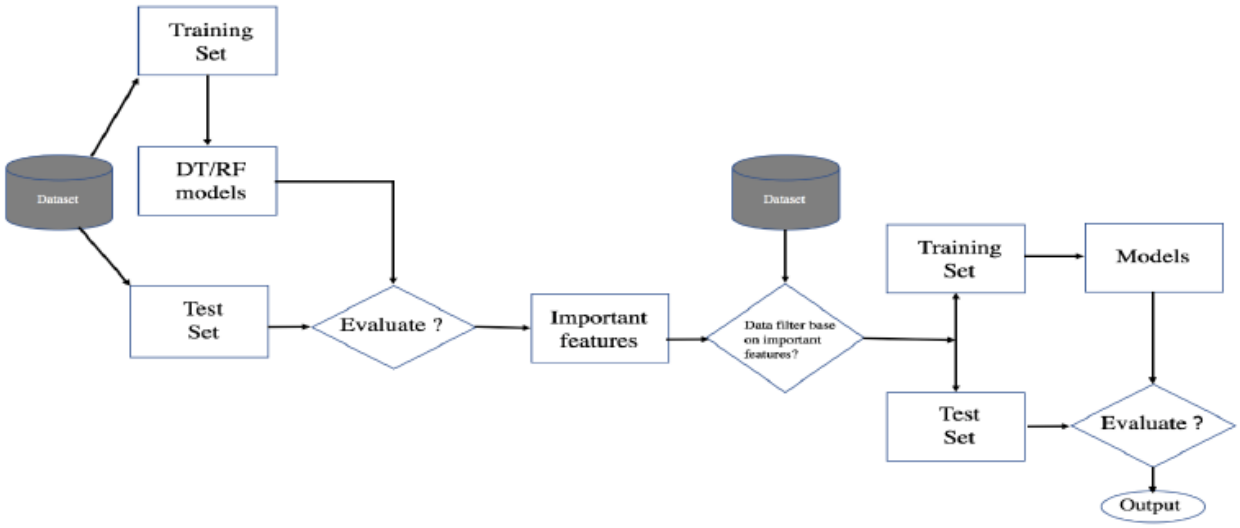

Figure 1. Transfer Learning Architecture

3 RESULTS AND FINDINGS

- We created models using classification algorithms Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN) and Multilayer Perceptron (MLP) using the initial dataset containing over 50 features. We found the best model for each algorithm.
- We created models for each algorithm using transfer learning technique. Each model was trained using only important features identified in the previous step.
- We also used important features identified by an expert to train and test models.
- Finally, we combined the sets of important features identified by classification algorithms with those identified by the expert and repeated model building and evaluation.

In the following sections, we present a comprehensive comparison of the best models of these techniques.

3.1 The Dataset

We used the dataset from [11]. It contains records of 70,000 patient with one or more hospital visits for each patient for a total of 101,766 hospitalization records. In [11] the goal was to analyze re-hospitalization for the patients. We are using the data to determine the likelihood of diabetes patients developing heart conditions as well. The dataset lists 55 features for 101,766 hospitalization records. The data is available online as supplementary material at: http://dx.doi.org/10.1155/2014/781670.

3.2 Findings

Algorithm	Accuracy		Precision Micro		Precision Macro		Precision Weighted		Recall Micro		Recall Macro		Recall Weighted	
	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
Decission Tree	max_depth: 7	0.998435055	max_depth: 7	0.99838289	max_depth: 6	0.996739538	max_depth: 15	0.998287292	max_depth: 9	0.998435055	max_depth: 16	0.997467173	max_depth: 8	0.998330725
Random Forest	max_depth: 18	0.971465832	max_depth: 18	0.971465832	max_depth: 18	0.980816173	max_depth: 18	0.972169671	max_depth: 18	0.971465832	max_depth: 18	0.905284003	max_depth: 18	0.971465832
KNN	n_neighbors: 3	0.952164841	n_neighbors: 3	0.952164841	n_neighbors: 2	0.935128792	n_neighbors: 3	0.951244556	n_neighbors: 3	0.952164841	n_neighbors: 1	0.9009091	n_neighbors: 3	0.952164841
MLP	max_iter: 90000	0.922222222	max_iter: 110000	0.921804903	max_iter: 40000	0.899345017	max_iter: 160000	0.920852791	max_iter: 130000	0.923682838	max_iter: 200000	0.800657107	max_iter: 190000	0.923109025

Figure 2. Best Models using all features

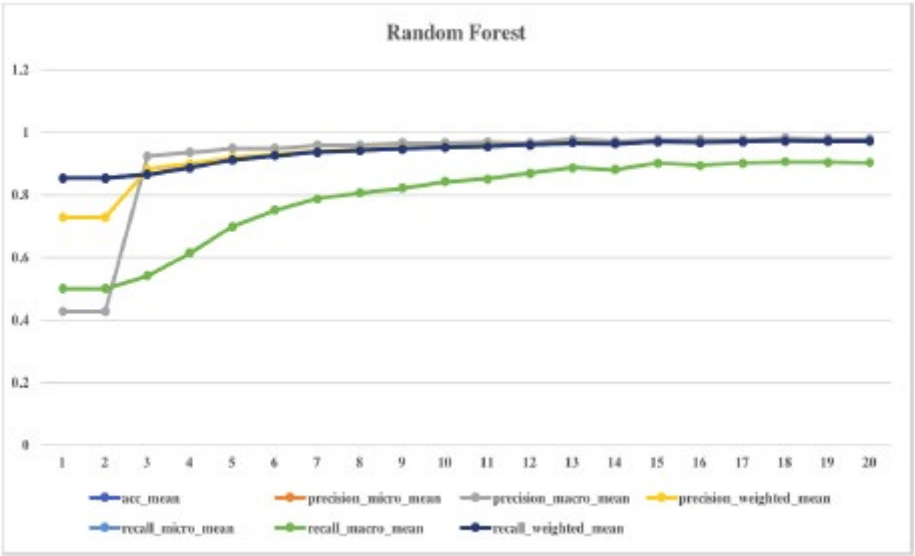

Figure 3. Line graph of random forest with varying max_depth using all features.

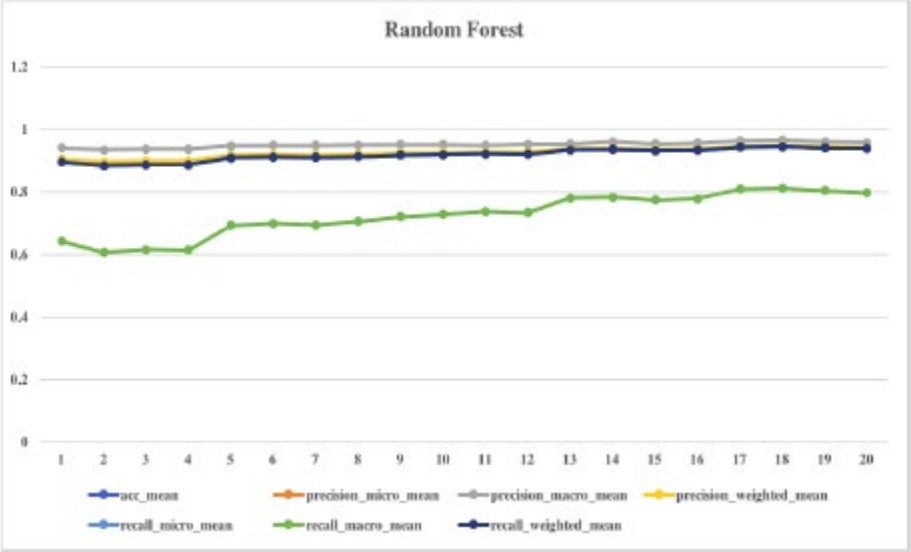

Figure 4. Line graph of random forest with varying max_depth and min_sample_split using all features.

3.2.1 Using all the features of the dataset.

We trained each model using all the features available in the dataset, and then we evaluated each model by calculating evaluation matrixes using grid search with 5-fold cross validation. The best models of the Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN) and Multilayer Perceptron (MLP) algorithms are shown in Figure 2. To save space, we show the line graphs only for the Random Forest algorithm (Figures 3 and 4).

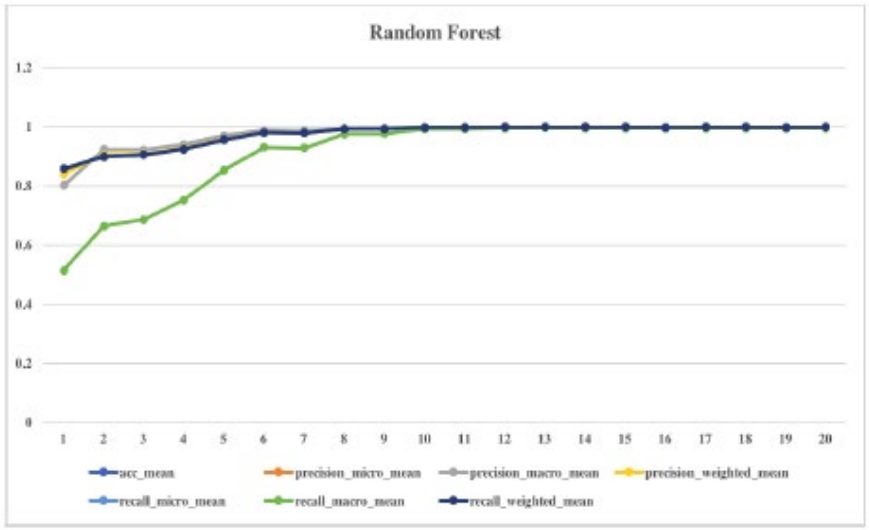

Figure 5. Line graph of random forest with varying max_depth using transfer learning.

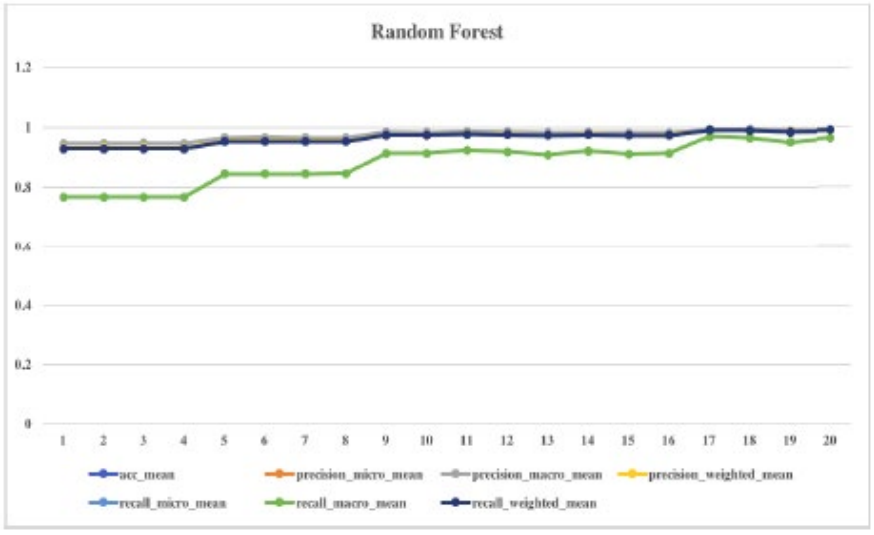

Figure 6. Line graph of random forest with varying max_depth and min_sample_split using transfer learning.

3.2.2 Using Transfer learning.

In transfer learning technique, for heart disease dataset we identified the top 10 important features during transfer learning using decision tree. And these top 10 features were only used for training all the models of Decision Tree, Random Forest, K-Nearest Neighbor and Multilayer Perceptron algorithms to demonstrate the transfer learning in this experiment. For estimating the evaluation matrixes, we also used the grid search with 5-fold cross validation and compared the models. To save space, we show the line graphs only for the Random Forest algorithm (Figures 5 and 6).

The following table (Figure 7) shows the best models of Decision Tree, Random Forest, K-Nearest Neighbor and MLP algorithms for each evaluation matrix.

Algorithm	Accuracy		Precision Micro		Precision Macro		Precision Weighted		Recall Micro		Recall Macro		Recall Weighted	
	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
Decission Tree	max_depth: 18	0.999113198	max_depth: 7	0.999061033	max_depth: 11	0.998168142	max_depth: 9	0.999115951	max_depth: 7	0.999061033	max_depth: 9	0.998305135	max_depth: 13	0.999113198
Random Forest	max_depth: 13	0.998539384	max_depth: 13	0.998539384	max_depth: 13	0.998552782	max_depth: 13	0.998539643	max_depth: 13	0.998539384	max_depth: 13	0.995619396	max_depth: 13	0.998539384
KNN	n_neighbors: 1	0.982994262	n_neighbors: 1	0.982994262	n_neighbors: 2	0.971366674	n_neighbors: 1	0.983024445	n_neighbors: 1	0.982994262	n_neighbors: 1	0.966387175	n_neighbors: 1	0.982994262
MLP	max_iter: 140000	0.921178925	max_iter: 90000	0.920918101	max_iter: 20000	0.902587897	max_iter: 170000	0.919562291	max_iter: 160000	0.922065728	max_iter: 40000	0.790494144	max_iter: 110000	0.920344288

Figure 7. Best models for heart disease dataset using transfer learning.

3.2.3 Using Expert-Identified Features.

For this experiment, experts identified 11 important features out of heart disease dataset as shown in section 3.3. We used these identified important features for training all the models of Decision Tree, Random Forest, K-Nearest Neighbor and MLP algorithm and estimated the evaluation matrix for each model using grid search 5-fold cross validation. To save space, we show the line graphs only for the Random Forest algorithm (Figures 8 and 9).

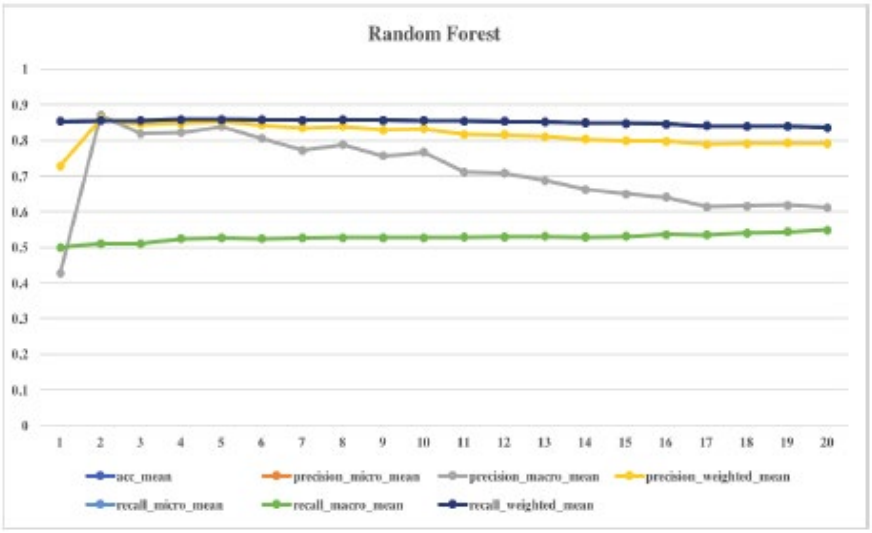

Figure 8. Line graph of decision tree with varying max_depth using features identified by experts.

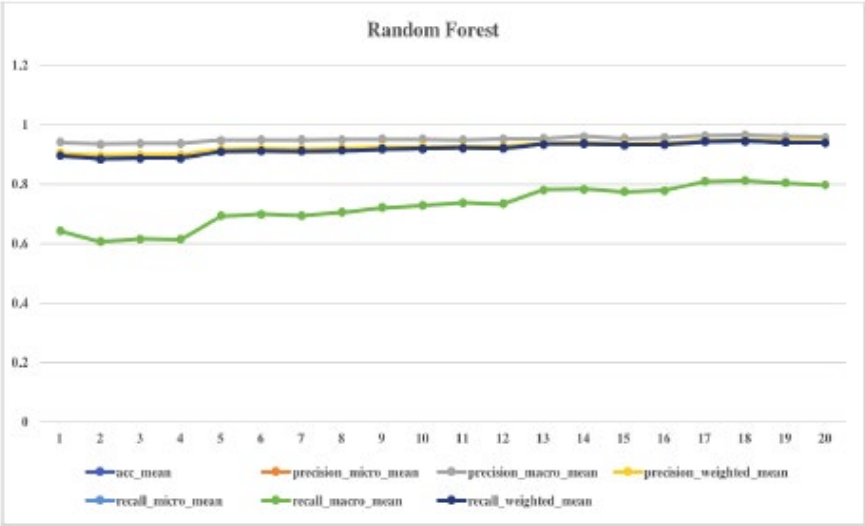

Figure 9. Line graph of random forest with varying max_depth and min_sample_split using features identified by experts.

The table in Figure 10 shows the best models of Decision Tree, Random Forest, K-Nearest Neighbor and MLP algorithms for each evaluation matrix.

Algorithm	Accuracy		Precision Micro		Precision Macro		Precision Weighted		Recall Micro		Recall Macro		Recall Weighted	
	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
Decission Tree	max_depth: 4	0.858581116	max_depth: 4	0.858581116	max_depth: 4	0.837718605	max_depth: 4	0.852828512	max_depth: 4	0.858581116	max_depth: 18	0.543696941	max_depth: 4	0.858581116
Random Forest	max_depth: 5	0.8590506	max_depth: 5	0.8590506	max_depth: 2	0.86990325	max_depth: 2	0.859760048	max_depth: 5	0.8590506	max_depth: 20	0.548233019	max_depth: 5	0.8590506
KNN	n_neighbors: 14	0.85722483	n_neighbors: 14	0.85722483	n_neighbors: 20	0.748469484	n_neighbors: 20	0.827417288	n_neighbors: 14	0.85722483	n_neighbors: 1	0.554452224	n_neighbors: 14	0.85722483
MLP	max_iter: 150000	0.857642149	max_iter: 190000	0.857746479	max_iter: 30000	0.816256373	max_iter: 160000	0.844376754	max_iter: 20000	0.857850809	max_iter: 130000	0.52766627	max_iter: 100000	0.857902973

Figure 10. Best models for heart disease dataset using features identified by experts.

3.2.4 Combining Transfer Learning and Expert-Identified Features.

For our final experiment, we combined the top 10 features identified during transfer learning (Section 3.2.2) with important features of heart disease dataset identified by experts (Section 3.2.3). We used these combined features for training all the models of Decision Tree, Random Forest, K-Nearest Neighbor and MLP algorithms. and estimated the evaluation matrix for each model using grid search 5-fold cross-validation. For this experiment, experts identified 11 important features To save space, we show the line graphs only for the Random Forest algorithm (Figures 11 and 12).

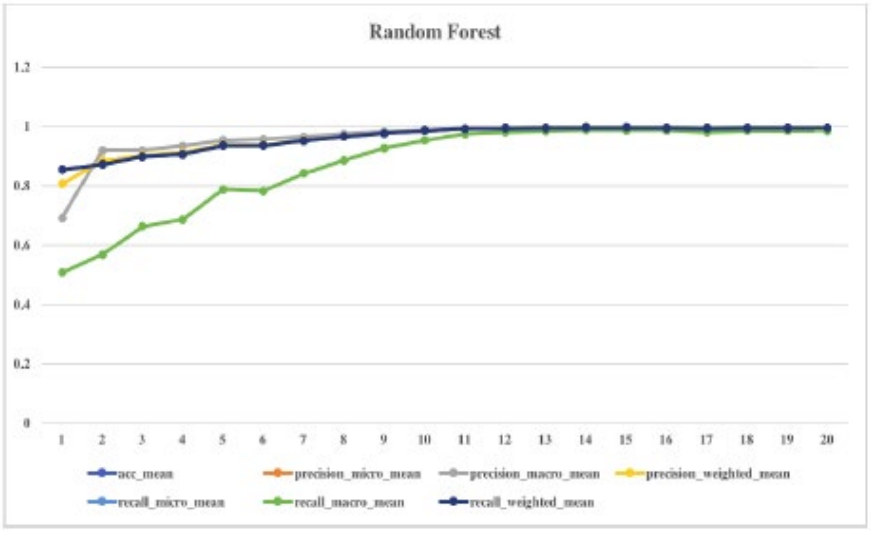

Figure 11. Line graph of decision tree with varying max_depth using transfer learning combined with features identified by experts.

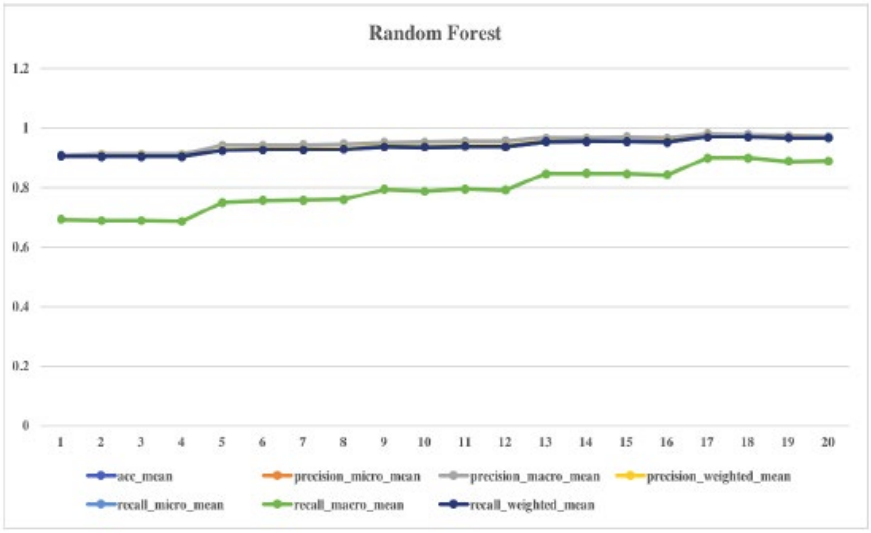

Figure 12. Line graph of random forest with varying max_depth and min_sample_split using transfer learning combined with features identified by experts.

Algorithm	Accuracy		Precision Micro		Precision Macro		Precision Weighted		Recall Micro		Recall Macro		Recall Weighted	
	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
Decission Tree	max_depth: 8	0.999061033	max_depth: 11	0.999113198	max_depth: 20	0.998345457	max_depth: 15	0.999168132	max_depth: 12	0.999061033	max_depth: 12	0.998305143	max_depth: 12	0.999165363
Random Forest	max_depth: 14	0.995774648	max_depth: 14	0.995774648	max_depth: 15	0.995000378	max_depth: 14	0.995768866	max_depth: 14	0.995774648	max_depth: 14	0.988859848	max_depth: 14	0.995774648
KNN	n_neighbors: 1	0.983098592	n_neighbors: 1	0.983098592	n_neighbors: 2	0.972378469	n_neighbors: 1	0.983179918	n_neighbors: 1	0.983098592	n_neighbors: 1	0.967623898	n_neighbors: 1	0.983098592
MLP	max_iter: 80000	0.924621805	max_iter: 120000	0.923943662	max_iter: 17000	0.903664631	max_iter: 110000	0.920105396	max_iter: 190000	0.923735003	max_iter: 20000	0.802112184	max_iter: 30000	0.925299948

Figure 13. Best models for heart disease dataset using transfer learning combined with features identified by experts.

The table in Figure 13 shows the best model of Decision Tree, Random Forest, K-Nearest Neighbor and MLP algorithms for each evaluation matrix.

3.3 Comparison

In this section, we compare all the 4-methodologies used with heart disease dataset, following table shows the comparisons between best models created by modifying one coefficient for each methodology and each machine learning algorithms.

The table shows the transfer learning methodology has better or almost same accuracy for all the machine learning algorithms comparing to other methodology. Here it also shows that the Decision Tree algorithm outperformed to be best among all the machine learning algorithm for all the methodology. It also shows that the model trained with only expert suggested features has the lowest accuracy for each algorithm. It also concludes that if the suggested features are combined with transfer learning, it outperformed to be best among all the methodology.

Features used for training	DT	RF	KNN	MLP
All 53 features	0.998	0.971	0.952	0.922
Transfer learning with top 10 features	0.999	0.998	0.982	0.921
Features identified by experts	0.858	0.859	0.857	0.858
Combined top 10 transfer learning features and expert-identified features	0.999	0.995	0.983	0.924

Figure 14. Comparison of the best models.

4 CONCLUSION

We compared the performance of the Decision Tree, Random Forest, K-Nearest Neighbor and Multilayer Perceptron machine learning algorithms. We chose an application where hospitalization data about diabetes patients are used to predict those that are likely to develop heart condition as well. We were, in particular, interested in assessing the improvements obtained by the inductive transfer learning approach. In traditional approach where all available features of the dataset were used to train the model, the best model of the Decision Tree outperformed with an accuracy of 99.84% for heart disease dataset, followed by Random Forest, KNN and MLP.

In the transfer learning technique, the top ten important features were identified out of all features using Decision Tree, and these important features were used to train the models. This technique showed that all algorithms performed the same or better compared to the traditional approach. Here too the best model of the Decision Tree classification algorithm outperformed the others with an accuracy of 99.91%.

We also performed experiments with expert-suggested features, which showed the performance of the model dropped compared to the transfer learning approach.

Finally, we experimented with transfer learning using a combination of expert-suggested and algorithm-generated top features. The best model of the Decision Tree algorithm had the

accuracy of 99.91% in this case, similar to the transfer learning without expert suggested features.

We conclude that inductive transfer learning can be used for feature reduction, and, it can also improve the accuracy of the prediction models.

REFERENCES

[1] Aurélien Géron. 2017. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. "O'Reilly Media, Inc.".

[2] Norberto A. Goussies, Sebastián Ubalde, and Marta Mejail. 2014. Transfer learning decision forests for gesture recognition. *Journal of Machine Learning Research* 15, 1 (2014), 3667–3690. http://dl.acm.org/citation.cfm?id=2750362

[3] Erico Guizzo. 2011. How Googles self-driving car works. IEEE Spectrum Online.

[4] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 46, 1-3 (2002), 389–422. https://doi.org/10.1023/A:1012487302797

[5] Toshihiro Kamishima, Masahiro Hamasaki, and Shotaro Akaho. 2009. TrBagg: A Simple Transfer Learning Method and its Application to Personalization in Collaborative Tagging. In *ICDM 2009, The Ninth IEEE International Conference on Data Mining, Miami, Florida, USA, 6-9 December 2009*. 219–228. https://doi.org/10.1109/ICDM.2009.9

[6] Pat Langley and Wayne Iba. 1993. Average-Case Analysis of a Nearest Neighbor Algorithm. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambéry, France, August 28 - September 3, 1993*. 889–894. http://ijcai.org/Proceedings/93-2/Papers/008.pdf

[7] Zhongqi Lu, Weike Pan, Evan Wei Xiang, Qiang Yang, Lili Zhao, and Erheng Zhong. 2013. Selective Transfer Learning for Cross Domain Recommendation. In *Proceedings of the 13th SIAM International Conference on Data Mining, May 2-4, 2013. Austin, Texas, USA*. 641–649. https://doi.org/10.1137/1.9781611972832.71

[8] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng*. 22, 10 (2010), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[9] Margaret A Shipp, Ken N Ross, Pablo Tamayo, Andrew P Weng, and Jeffery L Kutok. 2002. Diffuse large b-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. Nature Medicine. , 68-74 pages.

[10] Jyoti Soni, Ujma Ansar, Dipesh Sharma, and Sunita Soni. 2011. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications* 17, 8 (2011), 43–48.

[11] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. 2014. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International* (2014). http://dx.doi.org/10.1155/2014/781670

[12] Robert R. Trippi and Efraim Turban (Eds.). 1992. *Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real World Performance*. McGraw-Hill, Inc., New York, NY, USA.

[13] Ian H. Witten, Frank Eibe, and Mark A. Hall. 2011. *Data mining: practical machine learning tools and techniques, 3rd Edition*. Morgan Kaufmann, Elsevier. http://www.worldcat.org/oclc/262433473

[14] Yi Yao and Gianfranco Doretto. 2010. Boosting for transfer learning with multiple sources. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*. 1855–1862. https://doi.org/10.1109/CVPR.2010.5539857