Variant component principal linear reduction for prediction of hypothyroid disease using machine learning

M. Shyamala Devi¹, P. Swathi², Ayesha Jahangir³, A.Dhanunjai Sai Kumar⁴, A.Naga Bhushan Naidu⁵, M.Raghuram Manas⁶

¹ Professor, Computer Science & Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai.
², ³ Final Year B.Tech Student, Computer Science & Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai.
⁴, ⁵, ⁶ Third Year B.Tech Students, Computer Science & Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai.
shyamalapmr@gmail.com

Abstract. With the tremendous technological growth, the world is shifted to adapt the different food and life style by the people that results in the improper working of the body organs. The change in the food habits leads to a major problems that we face in the current scenario is the presence of hypothyroid in the body. The likelihood of hypothyroid still ruins as a challenging issue due to the uncertainty of proper symptoms. With this background, the machine learning can be used towards health care scenarios for the prediction of disease based on the patients past history. This paper focus on predicting the existence of hypothyroid with respect to the patients’ medical parameters. The hypothyroid patient dataset is taken from the UCI Metadata repository with 24 columns and 3163 unique patient’s records is used for the experimentation of hypothyroid with the following contributions. Firstly, the hypothyroid dataset from UCI machine repository is subjected with the data processing and exploratory analysis of the dataset. Secondly, the unrefined data set is fixed with different classifier algorithm to find the presence of hypothyroid and to examine the efficiency metrics before and after feature scaling. Thirdly, the data is processed to PCA with various combination of components as 5, 7 and 10 and is fixed with different classifier algorithm to examine the efficiency metrics before and after feature scaling. Fourth, the data is processed to LDA with various combination of components as 5, 7 and 10 and is fixed with different classifier algorithm to examine the efficiency metrics before and after feature scaling. Experimental results show that the Kernel Support Vector Machine classifier is found to have the accuracy of 99.52% for all the 10, 7, 5 component reduced PCA dataset. Similarly, the Logistic Regression, Kernel Support Vector Machine and Gaussian Naïve Bayes classifier is found to have the accuracy of 99.52% for all the 10, 7, 5 component reduced LDA dataset.

1. Introduction
The Hypothyroid disease is the underperformance situation in which the thyroid gland fails to secrete the needed quantity of the important hormones in the body. The hypothyroid is the difficult infectious disease that causes due to the unintended the thyroid simulating hormones and it level of existence in the body. It is also caused due to the constructive problem in the growth of the thyroid gland itself. The hypothyroid can be stabilized with the safe situation where the body has to stimulate the
antibodies that fights with the unwanted hormones. The machine learning technology were used for health care application in the recent days for the prediction of diseases.

2. Literature review
The paper proposes with the experiment that was done on 3710 instances and 29 features of thyroid patients. The entire future prediction of the hypothyroid disease is categorized as two classes with positive and negative class. The efficiency of the outcome was implemented by making different combinations of clinical values and it is classified with the thyroid dataset [1]. This paper makes an assumption that the prediction of hypothyroid can be done with the logistic regression, decision trees for the improvement in the performance of the prediction [2]. The comparative study on thyroid disease diagnosis by using Support Vector Machine, Multiple Linear Regression, Naïve Bayes and Decision Trees. The results were compared and it was seen that Decision Trees could be successfully used to help the diagnosis of thyroid disease [3]. An attempt is made to analyze Logistic regression and Support Vector Machine for multiclass classification of thyroid dataset [4].

The Artificial Neural network models can be used for the functioning of the thyroid gland networks and it is fed into the various input levels for the prediction of the hypothyroid gland. The resultant outcome of the thyroid disease prediction can be highly improved by including the biological process that is related to the contribution of thyroid gland to the disease [5]. The simulation results show that One-Against-All Support Vector Machines are superior to One-Against-One Support Vector Machines with polynomial kernels. The accuracy of OAASVM is also higher than AdaBoost and Decision Tree classifier on hypothyroid disease datasets from UCI machine learning dataset [6]. The CAD system is comprised of three stages. Focusing on dimension reduction, the first stage applies PCA to construct the most discriminative new feature set [7]. The proposed system in this paper is used for system switches to the second stage whose target is model construction. ELM classifier is explored to train an optimal predictive model whose parameters are optimized [8].

In this research work, various classification models are used to classify thyroid disease based on the parameters like TSH, T4U and goiter [9]. The objective of this research paper observation is to determine hyperthyroidism, hypothyroidism and thyroids participation in hormones can be good predictor of the final result of laboratories and to examination whether the propose ensemble approach can be similar accuracy to other single classification algorithm [10]. In this paper, it presents an enhanced fuzzy knearest neighbor (FKNN) classifier based computer aided diagnostic system for thyroid disease [11]. This paper proposes a framework Thyroid Disease Types Diagnostics that aims to assist the physicians during the diagnostic process of thyroid diseases in a very structured and transparent manner [12]. The purpose of this study is predication of thyroid disease using different classification techniques and also to find the TSH, T3, T4 correlation towards hyperthyroidism and hypothyroidism and also to finding the TSH, T3, T4 correlation with gender towards hyperthyroidism and hypothyroidism [13]. The performances of the classification algorithms have been analyzed on breast cancer and hypothyroid datasets [14]. The proposed model is to categorize the usage of the thyroid gland to the secretion of the hormones with the help of feature reduction and classification methods [15].

3. Our contributions
The main analysis of this paper is to analyze how well the classification algorithms are predicting and maintaining the accuracy level with the change in the component reduction with principal component and linear Discriminant analysis. The overall architecture of this paper is shown in figure.1. The existence of hypothyroid is predicted using machine learning classification algorithms with the following contributions.

(i) Firstly, the hypothyroid dataset from UCI machine repository is subjected with the data processing and exploratory analysis of the dataset.

(ii) Secondly, the unrefined data set is fixed with different classifier algorithm to find the presence of hypothyroid and to examine the efficiency metrics before and after feature scaling.
(iii) Thirdly, the data is processed to PCA with various combination of components as 5, 7 and 10 and is fixed with different classifier algorithm to examine the efficiency metrics before and after feature scaling.

(iv) Fourth, the data is processed to LDA with various combination of components as 5, 7 and 10 and is fixed with different classifier algorithm to examine the efficiency metrics before and after feature scaling.

4. Implementation Setup

4.1. Dataset Exploratory Analysis

The python scripting language is coded in Spyder editor with Anaconda navigator for implementation. The dataset information is shown in figure. 2. The feature analysis of hypothyroid dataset is depicted as density plot and coorelation matrix is shown in figure. 3 - figure. 6.

Figure 1. Overall proposed workflow

Figure 2. Dataset information

Figure 3. Dataset density plot
5. Results and discussions
The hypothyroid dataset from UCI machine repository is subjected with the data processing and exploratory analysis of the dataset. The unrefined data set is fixed with different classifier algorithm to find the presence of hypothyroid and to examine the efficiency metrics before and after feature scaling and is shown in figure. 7–figure.8. The data is processed to PCA with various combination of components as 5, 7 and 10 and is fixed with different classifier algorithm to examine the efficiency metrics before and after feature scaling. The cumulative variance VS Evariance of the 10 component PCA is shown in figure 9 – figure 10. The classifier Performance analysis of the 10 component PCA before and after feature scaling is and is shown in figure. 11– figure 12.
Figure 7. Performance metrics for raw dataset before feature scaling.

Figure 8. Performance metrics for raw dataset after feature scaling.

Figure 9. Non Scaled CVariance VS Evariance for PCA with 10 components.

Figure 10. Scaled CVariance VS Evariance for PCA with 10 components.
Figure 11. Performance metrics for 10 component pca before feature scaling.

Figure 12. Performance metrics for 10 component pca after feature scaling.

The cumulative variance VS Evariance of the 7 component PCA is shown in figure 13 – figure 14. The classifier Performance analysis of the 7 component PCA before and after feature scaling and is shown in figure. 15– figure 16. The cumulative variance VS Evariance of the 5 component PCA is shown in figure 17 – figure 18. The classifier Performance analysis of the 5 component PCA before and after feature scaling and is shown in figure. 19 – figure 20.
Figure 13. Non scaled CVariance VS Evariance for PCA with 7 components.

Figure 14. Scaled CVariance VS Evariance for PCA with 7 components.

Figure 15. Performance metrics for 7 component PCA before feature scaling.

Figure 16. Performance metrics for 7 component PCA after feature scaling.
Figure 17. Non-scaled CVariance VS Evariance for PCA with 7 components.

Figure 18. Scaled CVariance VS Evariance for PCA with 7 components.

Figure 19. Performance metrics for 5 component PCA before feature scaling.

Figure 20. Performance metrics for 5 component PCA after feature scaling.
The data is processed to LDA with various combination of components as 5, 7 and 10 and is fixed with different classifier algorithm to examine the efficiency metrics before and after feature scaling. The classifier Performance analysis of 10 component LDA before and after feature scaling is shown in figure 21 – figure 22.

Figure 21. Performance metrics for 10 component LDA before feature scaling.

Figure 22. Performance metrics for 10 component LDA after feature scaling.

The classifier Performance analysis of the 7 component LDA before and after feature scaling and is shown in figure. 23 – figure 24.
Figure 23. Performance metrics for 7 component LDA before feature scaling.

| Index | Classifier          | Precision | Recall | FScore | Accuracy |
|-------|---------------------|-----------|--------|--------|----------|
| 0     | LogisticRegression  | 0.995285  | 0.995261 | 0.995192 | 0.995261 |
| 1     | KNeighborsClassifier| 0.995285  | 0.995261 | 0.995192 | 0.995261 |
| 2     | Kernel SVM Classifier| 0.995285  | 0.995261 | 0.995192 | 0.995261 |
| 3     | Gaussian Naive Bayes| 0.995285  | 0.995261 | 0.995192 | 0.995261 |
| 4     | Decision Tree Classifier| 0.99363  | 0.993681 | 0.99362 | 0.993681 |
| 5     | Extra Tree Classifier| 0.998521  | 0.998521 | 0.998521 | 0.998521 |
| 6     | Random Forest Classifier| 0.99363  | 0.993681 | 0.99362 | 0.993681 |
| 7     | Gradient Boosting Classifier| 0.99363  | 0.993681 | 0.99362 | 0.993681 |
| 8     | AdaBoost Classifier | 0.995285  | 0.995261 | 0.995192 | 0.995261 |
| 9     | Ridge Classifier    | 0.995285  | 0.995261 | 0.995192 | 0.995261 |
| 10    | Ridge ClassifierCV  | 0.995285  | 0.995261 | 0.995192 | 0.995261 |
| 11    | SGD Classifier      | 0.995285  | 0.995261 | 0.995192 | 0.995261 |
| 12    | Passive Aggressive Classifier| 0.995285  | 0.995261 | 0.995192 | 0.995261 |
| 13    | BaggingClassifier   | 0.99363  | 0.993681 | 0.99362 | 0.993681 |

Figure 24. Performance metrics for 7 component LDA after feature scaling.

The classifier Performance analysis of the 5 component LDA before and after feature scaling and is shown in figure. 25 – figure 26.
6. Conclusion
An attempt is made in this paper to analyze how well the classification algorithms are predicting and maintaining the accuracy level with the change in the component reduction with principal component and linear Discriminant analysis. Experimental results show that the Kernel Support Vector Machine classifier is found to have the accuracy of 99.52% for all the 10, 7, 5 component reduced PCA dataset. Similarly, the Logistic Regression, Kernel Support Vector Machine and Gaussian Naive Bayes classifier is found to have the accuracy of 99.52% for all the 10, 7, 5 component reduced LDA dataset.
References

[1] Tyagi A Mehra R and Saxena A 2018 Interactive Thyroid Disease Prediction System Using Machine Learning Technique Proc. IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC-2018) p 20-22

[2] Chaubey G Bisen D Arjaria S and Yadav V 2020 Thyroid Disease Prediction Using Machine Learning Approaches, https://doi.org/10.1007/s40009-020-00979z.

[3] Razia S Swathi P Krishna N and Sumana N 2018 A Comparative study of machine learning algorithms on thyroid disease prediction (International Journal of Engineering and Technology) p 315-319

[4] Godara S and Kumar S 2018 Prediction of Thyroid Disease Using Machine Learning Techniques (International Journal of Electronics Engineering) vol 10 p 787-793

[5] Razia S and Rao M 2016 Machine Learning Techniques for Thyroid Disease Diagnosis - A Review (Indian Journal of Science and Technology) vol 9 pp 28

[6] Chamasemani F and Singh Y 2011 Multi-class Support Vector Machine (SVM) classifiers – An Application in Hypothyroid detection and Classification. P 351-356

[7] Li L N Ouyang J H Chen H L and Liu D Y 2012 A computer aided diagnosis system for thyroid disease using extreme learning machine (Journal of Medical Systems) vol 36(5) p 3327-3337

[8] Raisinghani S Shamdasani R Motwani M Bahreja A and Raghavan Nair Lalitha P 2019 Thyroid Prediction Using Machine Learning Techniques (Advances in Computing and Data Sciences) p 140–150

[9] Chandel K Kunwar V Sabitha S Choudhury T and Mukherjee S 2016 A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques (CSI Transactions on ICT) vol 4(2-4) p 313–319

[10] Yadav Dhyan and Pal Saurabh 2019 Decision Tree Ensemble Techniques To Predict Thyroid Disease. Vol 8 p. 8242-8246.

[11] Liu D Y Chen H L Yang B L et al 2011 Design of an Enhanced Fuzzy k-nearest Neighbor Classifier Based Computer Aided Diagnostic System for Thyroid Disease (Journal of Medical Systems) vol 36(5) p 3243–3254

[12] Ahmed J and Soomrani M 2016 TDTS: Thyroid Disease Type Diagnostics p 44-50

[13] Begum A and Parkavi A 2019 Prediction of thyroid Disease Using Data Mining Techniques Proc. Int. Conf. on Advanced Computing & Communication Systems p 342-345

[14] Verma D and Mishra N 2017 Comparative analysis of breast cancer and hypothyroid dataset using data mining classification techniques Proc. Int. Conf. on Power, Control, Signals and Instrumentation Engineering p 1624-1626

[15] Shankar K Lakshmanaprabu S K and Gupta D 2020 Optimal feature-based multi-kernel SVM approach for thyroid disease classification (Journal of Supercomputing) vol 76 p 1128–1143