Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Design & Development of System for Early Detection of Thyroid Disease Using Machine Learning: A Review Paper

Sachin K. Latad[1], Prof. (Dr.) S. V. Pattalwar[2]

[1]Department of Electronics & Telecommunication Engineering
[2]Prof. Ram Meghe Institute of Technology & Research, Badnera, Dist. Amravati

Abstract

Thyroid problems are caused by an irregular and unnatural development of thyroid tissue at the edges about the gland that produces thyroid. Hyperthyroidism, or an overactive thyroid gland, and hypothyroidism, or an inactive thyroid gland, are the two main types of thyroid disorders. This research suggests using effective classifiers with or without methods for feature selection by employing machine learning algorithms to identify and diagnose thyroid disease, as measured by accuracy and other performance evaluation metrics. Thyroid ultrasound image data is available on the Neural Information Processing Systems (NIPS), The Challenge of Feature Selection as well as on keggle, Github website.

This research suggests three new training techniques for multilayer perceptron (MLP) binary classifiers, two of which are based on the method of back propagation and the third is about information theory. The principle of maximal-margin (MM) serves as foundation for both back propagation techniques. The third strategy provides a strong training framework that may utilize the most effective aspects of each suggested training technique. The basic idea is to use neurons taken from three different neural networks two based on MM and the third based on LM that have all been formerly received training from Levenberg Marquard (LM), MMGDX, and MICI, respectively, to create a neural model. This study includes features like body mass index (BMI), heart rate, and blood pressure because they are directly related to thyroid diseases and are essential for obtaining the most accurate results.

To determine the thyroid nodules' risk of malignancy, sonographic characteristics related to their edges, shape, size, and volume are utilized. An automated assessment of these properties would be made possible by automatically separating nodules from the normal thyroid gland.

IndexTerms – ASNN (Assembled Neural Network), Levenberg Marquard(LM), Multilayer Perceptron(MLP), Maximal Margin (MM).

1. Introduction

Neural networks can be viewed as a subfield of artificial intelligence, machine learning, statistics, parallel computing, and other disciplines. Neural networks are appealing because they excel at resolving issues that conventional computing techniques find most challenging. Examine a picture processing problem like identifying a commonplace object displayed against a background of other objects. Even the brain of a young child can complete this activity within a few tenths of an instant. However, it is quite difficult to create a traditional serial computer to perform as well. The same youngster, however, may not be able to compute 2+2=4, but the serial machine can do this in a matter of nanoseconds.

Simple mathematics is better handled serially, whereas the picture recognition problem is best solved in parallel. One important difference between the two issues is this. According to neurobiologists, the brain functions similarly to an analog computer with tremendous parallelism, with roughly 10^10 basic processors that take a few milliseconds to react to input. In certain real-world situations when it is extremely challenging to create a traditional algorithm, neural network technology allows us to apply parallel processing techniques.

The neural network algorithm takes inspiration from the way the brain analyzes information. The brain works incredibly well. With a response time that is almost 100,000 times slower than that of computer chips, it outperforms the computer in complex processes like motion control, image and sound recognition, and more. Additionally, it uses roughly 10,000,000,000 times less energy per operation than a computer chip. With 10^11 neurons, the brain is a multi-layer structure that functions as a parallel computer that can learn from the "feedback" it gets from the outside environment and creating new brain connections between neurons or adjusting the activity of already-existing ones to change its structure. In order to complete the picture, let's add note that a normal neuron is connected to between 50 and 100 other neurons.

2. Literature Review

Human brain research dates back thousands of years. It was only logical to attempt to use this thought process once sophisticated electronics were introduced. When young mathematician Walter Pitts and neurophysiologist Warren McCulloch authored a study on how neurons might function. The first step toward artificial neural networks was taken in 1943. Electrical circuits were utilized to simulate a basic neural network. Donald Hebb's 1949 book The Organization of Behavior made highlight the brain pathways are enhanced every time they're utilized, which is a key aspect of human learning.

The basic ideas of these theories about human cognition might be modeled when computers developed into their infancy in the 1950s. The initial attempt to model a neural network was initiated by Nathanial Rochester of IBM's research labs. The derivation of back-propagation in 1986 marked the beginning of the contemporary era of neural networks. In the ten years since parallel distributed processing was rewritten, (Rumelhart and McClelland, 1986) Neural networks have been the subject of a vast quantity of literature. Since neural networks are used in so many different fields, it is exceedingly challenging to take in all of the information that is available. To help readers to know the origins of neural network evolution, neural networks' brief history has been prepared. Additionally, a thorough analysis of the algorithm for back-propagation and feed-forward neural network has been written for this study. To demonstrate the requirement for the planned work in this study, papers on a variety of related topics are described in depth. Ten years isn't a long time for study, though, so no single book has established itself as the foremost knowledgeable about neural networks.

The back-propagation algorithm, Werbos's 1974 discovery was rediscovered in 1986 by Williams, Hinton, and Rumelhart in their book Learning Internal Representation by Error Propagation. Back-propagation is a gradient descent modification technique that artificial neural networks utilize for curve-fitting and reduction.

The annual international conference of IEEE for ANN researchers began in 1987. The Society for International Neural Networks (INNS) and INNS Neural Networking, its magazine, was founded in 1987 and 1988, respectively. "Beyond feed-forward models trained by back-propagation: A practical training tool for a more efficient universal approximator": MLP and simultaneous recurrent neural networks (SRNs), two of the numerous complex neural models, are both universal approximators. Any continuous, multi-dimensional, nonlinear function can be approximated by MLP with any level of precision.

MLP's simplicity makes it a popular model for applications using pattern recognition. The MLP is used to outperform (or at least perform comparably to) state-of-the-art methods like Support vector machines (SVMs), Bayesian neural networks, and nonlinear kernels, or innovative algorithms using kernel Fisher discriminant analysis as the basis.

3. RESEARCH GAPS

Authors	Paper Title	Proposed Methodology	Outcome	Other outcome	Research Gaps
"Pal, R.; Anand, T.; Dubey, S.K." [29]	"Evaluation and performance analysis of classification techniques for thyroid detection"	Naive Bayes, SVM, KNN	Accuracy is 94.70%, 92.70%, 96.90% for Naive Bayes, SVM, KNN resp., using KEEL repository thyroid disease dataset. KNN have Error rate without feature selection 8.61% with feature selection technique 2.16% and 3.23% respectively.	Recall without feature selection 90% with feature selection 96% and 95% respectively. Specificity without feature selection 95% with feature selection technique 98% and 98% respectively	It is predicted that newly added features will offer precise and reliable measurements for thyroid illness diagnosis.
"Chandel, K.; Kunwar, V.; Sabitha,	"Comparative study on thyroid disease detection using KNN and	KNN, Naive Bayes	Accuracy in percentage 93.44%, 22.56% for KNN, Naive Bayes respectively, using KEEL repository thyroid disease dataset.	Recall without feature selection 67% with feature selection 67% and 67% respectively. Specificity without feature selection 51% with feature selection technique 80% and 83% respectively	Newly incorporated features are thought to offer precise and reliable measurements for

S.; Chou dhur y, T.; Muk herje e, S" [30]	Naive Bayes classifica tion technique s"		DT have Error rate without feature selection 25.81% with feature selection technique 24.66% and 23.08% respectively.		thyroid illness diagnosis.
"Bek ar, E.T.; Ulut agay, G.; Kant arcı, S. " [31]	"disease by using data mining models: a comparis on of decision tree algorithm s"	NBTREE, LADTREE, REPTREE, BFTREE	Accuracy about 75.00%, 66.25% ,62.50% , 65.00 % for NBTREE, LADTREE, REPTREE, BFTREE respectively, using local hospital dataset. SVM have Error rate without feature selection 19.54% with feature selection technique 13.98% and 13.98% respectively.	Recall without feature selection 70% with feature selection 76% and 79% respectively. Specificity without feature selection 85% with feature selection technique 80% and 90% respectively	It is predicted that recently added features will offer precise and accurate measuremen ts for thyroid illness diagnosis.
"Cha lekar , P.; Shro ff, S.; Pise, S.; Pani cker, S.S". [32]	"Use of K-nearest neighbor in thyroid disease classifica tion"	KNN	Accuracy about 97.00% for KNN, using UCI repository thyroid disease dataset. Logistic regression have Error rate without feature selection 9.68% with feature selection technique 0% and 1.08% respectively.	Recall without feature selection 88% with feature selection 100% and 98% respectively. Specificity without feature selection 94% with feature selection technique 100% and 99% respectively	It is predicted that newly added features will offer precise and reliable measuremen ts for thyroid illness diagnosis.
"Tya gi, A.; Mehr a, R.; Saxe na,	"Interacti ve thyroid disease predictio n system using	KNN, DT	Accuracy in percentage 97.50%, 75.00% for KNN, DT respectively, using UCI repository thyroid disease dataset	Recall without feature selection 86% with feature selection 99% and 97% respectively. Specificity without feature selection 94% with feature selection technique 100% and	It is predicted that recently added capabilities would offer precise and

A"	machine learning technique . "			99% respectively	accurate measurements for thyroid diagnosis.
"R. Ilin, R. Kozma, and P. J. Werbos" [8]	"Beyond feedforward models trained by backpropagation"	Choosing every possible combination of the two activation functions, logsig and tansig, for the output and hidden layer neurons	Recognition accuracy begins to decline when overfitting causes the number of hidden neurons to surpass a particular threshold. i.e. accuracy start decreasing below 85% for number of hidden neuron exceeds than 8 i.e its optimum value	Accuracy with CSRN & MLP respectively for input size 5x5 80 ± 6 % 66 ± 10 % Input size 6x6 82 ± 6 % 65 ± 12 % Input size 7x7 88.5 ± 6 % 63 ± 12 %	Without feature selection technique i.e. instance of supervised learning is not consider
"S. Young and T. Downs" [3]	"CARVE : A constructive algorithm for real-valued examples "	Gorman and Sejnowski's CARVE experiments. A feed forward network is created by the algorithm using a single threshold unit hidden layer.	An algorithm establishes a feedforward network using a multiple concealed layer of threshold units	CARVE (25,65)Training, Testing set performance for network size 6.9±0.5, 100%, 81.2%±10.1% respectively CARVE-BP (25,65)Training, Testing set performance for network size 6.9±0.5, 95.3 %±0.3%, 84.2%±8.2% respectively CARVE second order for network size 8.1±0.6 Training, Testing set performance 100%, 76.9%±9.7% respectively	Without feature selection technique i.e. instance of supervised learning is not consider
"K. Hornik, M. Stinchcombe, and H. Whit	"Universal approximation of an unknown mapping and its derivatives using	One hidden layer and an activation function for the hidden layer are used in multilayer feed-forward networks that is suitably	Advancement in the understanding practical application of multilayer feed-forward networks for approximating unknown mappings and their derivatives, ultimately contribute to the development of	Currently, there is no theoretical assurance that multilayer feed forward In general, networks can simultaneously estimate the derivatives of an unknown mapping. .	Without feature selection technique i.e. instance of supervised learning is not consider

e " [7]	multilayer feed-forward networks "	smooth	more robust and efficient machine learning algorithms.		
"C. Liu et. Al. " [9]	"Gabor-based kernel PCA with fractional power polynomial models for face recognition"	Methods feature extraction include Principal component analysis (PCA), independent component analysis (ICA), linear discriminant analysis (LDA), and robust PCA	Advancements in face recognition technology, improving its accuracy, robustness, efficiency, and usability for various applications in security, surveillance, authentication, and human-computer interaction.	PCA using a Gabor-based kernel and fractional power approach Using 246 features, the accuracy of the polynomial model (d=0.6) for facial recognition is 99.5%.	Without feature selection technique i.e. instance of supervised learning is not consider
"A. Ruiz and P. E. Lopez-de-Teruel" [11]	"Nonlinear kernel-based statistical pattern analysis"	A number of generalized linear models, including the radial basis function (RBF) network and the potential functions technique, are included in neural models.	This could involve investigating strategies to handle noisy or incomplete data, Adversarial robustness (i.e. improve the security and reliability of the models)	SVM's exceptional generalization ability has been shown in both more challenging tasks (PIMA) and learning scenarios with minimal class overlap (MONK1, IONOSPHERE, MONK3). However, in the same problems, the other approaches can achieve extremely acceptable test error rates (e.g., KL and KM in MONK3, KM in IONO, or KL and KLR-10% in the difficult task PIMA) that are similar to SVM. In some situations, some techniques even perform better than SVM, particularly when	Without feature selection technique i.e. instance of supervised learning is not consider

				there is a large degree of class overlap (KM in MONK2, or KL and KLR-10% in ABALONE).	
"J. Yang, A. F. Frangi, J. U. Yang, D. Zhang, and Z. Jin" [12]	"KPCA plus LDA: A complete kernel Fisher discriminant framework for feature extraction and recognition"	Least square loss function and an ℓ2,1-norm regularization	Investigate transfer learning strategies to adapt the framework to specific biomedical domains or tasks with limited labeled data	Average recognition rate for polynomial function using GDA CKFD irregular CKFD regular CKFD LDA 81.87%±2.95%,84.48%±0.8%, 83.80%±0.62,86.96%±0.50, 76.34%±0.77% respectively. For Gaussian function using GDA CKFD irregular CKFD regular CKFD LDA 87.64%±2.35%, 88.59%±0.26%, 84.27%±0.58, 88.79%±0.31, 76.34%±0.77% respectively.	Without feature selection technique i.e. instance of supervised learning is not consider
"F. Ratle, G. Camps-Valls, and J. Weston" [13]	"Semi-supervised neural networks for efficient hyperspectral image classification"	Semi-supervised NN NN framework for classifying HSI on a wide scale ex. Eg. CNN, AE	Develop strategies to address class imbalance and rare class detection in hyperspectral image classification. i.e. inadequate generalization of the model and low classification accuracy	Results from the semisupervised part using the labeled pixels for the KSC picture (ten classes) under supervision. accuracy per class, OA (OA[%]) for batch k-means, online k-means, $NCutEmb^{max}$, $NCutEmb^{all}$ 67.42%± 0.64%, 70.75%± 0.67%, 77.41%± 0.75% 75.13%± 0.72% respectively.	Without feature selection technique i.e. instance of supervised learning is not consider
"T.M. Khoshgoftaar, J. Van Hulse, A. Napo	"Supervised neural network modeling empirical investigation into learning from imbalanc	The learning matrix and the linear associator, two well-known associative networks, are combined to create hybrid	Explore the use of active learning techniques to improve label quality in imbalanced datasets. Investigate methods for selecting the most informative instances for manual labeling, prioritizing instances	In actuality, MLP gains more from training data noise reduction than from CD balancing. This is not to suggest that MLP is exempt from the negative effects of class inequality. Limited model generalization and low classification accuracy, instead, the use of sampling typically	Further testing of MLP instead, the use of sampling typically does not mitigate these

litano" [14]	ed data with labeling errors"	associative memory. It provides thorough answers to a few problems with uneven data classification.	that are likely to be mislabeled or difficult for the model to classify accurately.	does not mitigate these effects.	effects.Without feature selection technique i.e. instance of supervised learning is not consider
"Jinghua Wang, Jane You, QinLi, YongXu" [15]	"Extract minimum positive and maximum negative features for imbalanced binary classification"	Finding the absolute values of the unbalanced binary classification using minimum positive and maximum negative features	In the context of imbalanced binary classification, explore techniques for visualizing and understanding the discriminative power of the highest negative and lowest favorable attributes, providing insights into the underlying data distribution.	It demonstrates how the suggested approach uses fewer feature extractors to achieve improved classification accuracy.	Without feature selection technique i.e. instance of supervised learning is not consider
"H. Peng, F. Long, and C. Ding" [16]	"Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy"	First derive an equivalent form, mRMR criterion, for the selection of first-order incremental features. Next, combine mRMR with additional advanced feature selectors (like wrappers) to demonstrate a feature selection procedure	Advancements in mutual information-based feature selection methods, enabling more effective and interpretable approaches for identifying relevant and non-redundant features in diverse datasets and application domains.	For HDR data set, classification error rates with classifiers NB, With SVM, the lowest mRMR and MaxRelis error rates are approximately 3.5 and 5.5 percent, respectively; with LDA, the lowest mRMR and MaxRel error rates are approximately 6 and 10 percent, respectively, the lowest MaxRel has an error rate of about 11%, while mRMR features have an error rate of about 7%.	Without appropriate feature selection technique i.e. instance of supervised learning is not consider.

		with two stages.			
"Simon G, Verleysen M." [22]	"High-dimensional delay selection for regression model with mutual information and distance-to-diagonal criteria"	Based on a geometrical heuristic, the distance-to-diagonal (DD) criterion, is another quick criterion that is suggested to minimize all delays for a reconstruction in a high-dimensional phase space.	Further, Using benchmark and fabricated time series, the criterion for distance to diagonal is demonstrated and contrasted with MI. In spaces larger than two dimensions, various delays are chosen using a k-nearest-neighbors (k-NN) algorithm-based high-dimensional mutual information estimator.	Negative readings are occasionally produced by larger dimensional MI. Because the estimator is known to be biased, The search for a minimum is not penalized by this circumstance. It is still evident from these final three tables that the high-dimensional MI and the DD criterion function similarly: No criterion is always superior to another; depending on the time series being utilized and the dimension being taken into consideration, there may be a benefit to either strategy.	Without appropriate feature selection technique i.e. instance of supervised learning is not consider
"Francois D, Rossi F, Wertz V, Verleysen M." [23]	"Resampling methods for parameter free and robust feature selection with mutual information"	This study suggests using the permutation test, resampling methods and a K-fold cross-validation to get around both issues.	The resampling methods reveal the variance of the estimator, which can be utilized to determine a threshold to halt the forward process and automatically set the parameter.	Combining For feature selection, Mutual information in conjunction with a forward process is a good option. Because it is nonlinear and nonparametric, it is in fact faster than a wrapper approach, which makes very few assumptions about the data and uses the prediction model for all assessments.	Estimation is frequently challenging in high-dimensional spaces, that is, when multiple features have already been chosen.

4. Proposed Methodology

Conventional method of thyroid disease diagnosis uses analysis of blood samples. The process based on image processing and supervised neural network gives more accuracy. The focus is to get higher accuracy and reduced error rate for thyroid data using machine learning algorithms.

The design and development of the machine learning algorithm to diagnose the thyroid disorders, process are divided into three algorithms. The first one is Maximum-Margin gradient descent with an adaptive learning rate algorithm (MMGDX). It stretches out the distance between value of the two classes (margin) to its maximum. It makes use of Area under ROC curve (AUC) information for the stopping criterion.

The second one is the Minimization of inter-class interference (MICI), which generates a hidden layer where patterns have desirable statistical distribution. It also uses AUC for stopping criterion over the validation data set. The last one is a novel algorithm named an Assembled neural network (ASNN) which composes hidden neurons from previously trained algorithms. The accuracy of ASNN will be compared with another state-of-art classifier.

The focus is on a non-linear classifier. In machine Classification is seen as an example of supervised learning, which is learning in which a training set of, correctly identified observations is available.

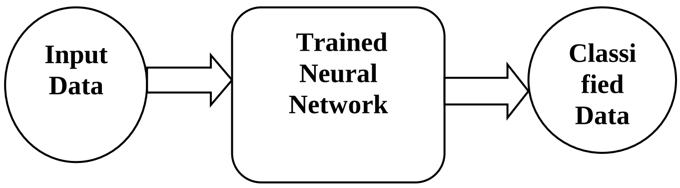

Figure 4.1: BASIC FRAMEWORK

Above figure shows basic framework of classifier. It is binary classifier that is non-linear for several kinds of application. Input data contains feature matrix and this classifier gives output by using trained neural network. For correct classification, it uses three novel training algorithms. The capability of Supervised assembled neural network model to generalize data depends on several factors such as The distribution of the input-output dataset and the proper choice of the system's input-output parameters and how the input-output dataset is presented to the neural network.

The training and testing data sets are available on Keggle website as well as Neural Information Processing Systems (NIPS), Feature Selection Challenge. Data sets of Thyroid have different characteristics, in the number of features, the margins, shape, size, and volume of thyroid. Their risk of malignancy is evaluated using nodules. Automatically segmenting nodules from normal thyroid gland would make it possible to estimate these properties automatically. Ultrasonic image data is used for thyroid images.

The flow of execution of these algorithms is shown in Figure 4.5.2 Name of these algorithms are listed below:

1. Maximum-Margin gradient descent with adaptive learning rate algorithm (MMGDX)
2. Minimization of inter-class interference (MICI)
3. Assembled neural network (ASNN)

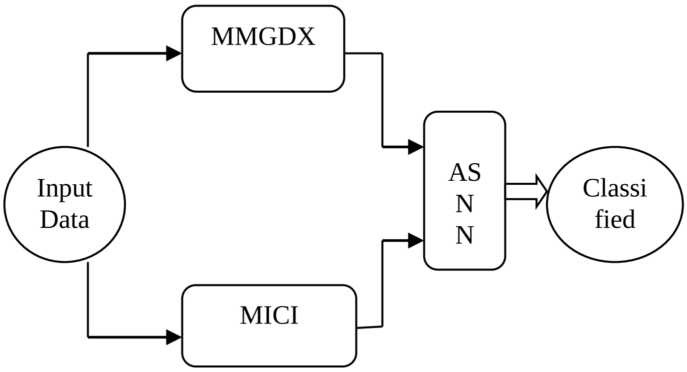

Figure 4.2: FLOW OF EXECUTION

Maximum-Margin gradient descent with adaptive learning rate (MMGDX)

The gradient descent error is equivalent to the desired output less the estimated output. But in MMGDX, the support vector norm, target output, and estimated output are used to calculate error. Information from the AUC curve is used for stopping criteria. The output and hidden layers of the MMGDX algorithm are optimized simultaneously in a single process. To generate an output that is hidden that is specifically directed toward a bigger margin for the output layer, Back-propagation of the objective function occurs between the output and hidden layers.

Minimization of interclass interference (MICI)

The second training method (MICI) has two phases, first one is hidden layer training and second one is output layer training. Hidden layer training is same as MMGDX, but Fisher linear discriminant analysis is used to train the output layer. It generates pattern-containing hidden layer with a favorable statistical distribution. MICI generates a hidden space where each class's patterns dispersion is reduced and the Euclidean distance between its prototypes is enhanced.

Assembled Neural Network (ASNN)

The third and the most important methodology named assembled neural network (ASNN), which uses neurons taken from previously trained networks, such as MICI and MMGDX, to create a neural model. But want to provide a strong training foundation. It is capable of utilizing any training method to its fullest. The goal of neurons is to maximize the information that is shared between the hidden and target outputs.

5. Conclusion

To obtain the most optimum values of accuracy and selection of robust neural networks for early detection of thyroid disorders, a modified machine learning-based ANN can be adopted in the future with continuous training. Generalization is the ability to train with one data set and then successfully classify independent test sets. Although continued training will increase the accuracy of the training set, it may cause the accuracy of the test set to decrease after a certain point. Over-fitting means the error is larger for testing data sets. Several performance comparisons between neural and conventional classifiers have been made by many studies, which will be considered at different stages of implementation to a proposed supervised neural network of thyroid disorder dataset.

One of the major advantages of neural networks is their ability to generalize. This means that a trained network could classify data from the same class as the learning data that it has never seen before. In real-world applications, developers normally have only a small part of all possible patterns for the generation of a neural network. To reach the best generalization, the dataset should be split into three parts:

- The training set is used to train a neural network. The error of this dataset will be minimized during training.
- The validation set will be used to determine the performance of a neural network on patterns that are not trained during learning.
- A test set will be established for finally checking the overall performance of a neural network.

The expected outcomes at the end of the implementation will be the comparison of various approaches and optimization based on their performance.

References

1. Oswaldo Ludwig and Urbano Nunes, "Novel Maximum-Margin Training Algorithms for Supervised Neural Networks," in IEEE Transactions on Neural Networks, vol. 21, no. 6, pp. 972-984, June 2010.
2. Shigeo Abe "Support Vector Machines for Pattern Classification," Springer Textbook, ISSN 1617-7916, 2010.
3. S. Young and T. Downs, "CARVE: A constructive algorithm for real-valued examples," IEEE Trans. Neural Network., vol. 9, no. 6, pp. 1180-1190, Nov. 1998.
4. T. Nishikawa and S. Abe, "Maximizing margins of multilayer neural networks," in Proc. 9th Int. Conf. Neural Inf. Process., Singapore, No 2002, vol. 1, pp. 322-326.
5. A. K. D. Jayadeva and S. Chandra, "Binary classification by SVM based tree type neural network," in Proc. Int. Conf. Neural Network, vol. 3, pp. 2773-2778, May 2002.
6. U. Franke and S. Heinrich, "Fast obstacle detection for urban traffic situations," IEEE Trans. Intell. Transp. Syst., vol. 3, no. 3, pp. 173-181, Sep. 2002.
7. K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," Neural Netw., vol.3, no. 5, pp. 551-560, 1990.
8. R. Ilin, R. Kozma, and P. J. Werbos, "Beyond feedforward models trained by backpropagation: A practical training tool for a more efficient universal approximator," IEEE Trans. Neural Network., vol. 19, no.6, pp. 929-937, June 2008.
9. C. Liu, "Gabor-based kernel PCA with fractional power polynomial models for face recognition," IEEE Trans.
10. Pattern Anal. Mach. Intell., vol. 26, no. 5, pp. 572-581, May 2004.