**Enhancing Ensemble Prediction Accuracy of Breast Cancer Survivability and Diabetes Diagnostic using optimized EKF-RBFN trained prototypes**

Vincent Adegoke1, Daqing Chen, Ebad Banissi and Safia Barsikzai

1corresponding author

Computer Science and Informatics, School of Engr’g, London South Bank University London, United Kingdom

adegokev@lsbu.ac.uk; chenq@lsbu.ac.uk; banisse@lsbu.ac.uk; barikas@lsbu.ac.uk

# **Abstract**

We are in a machine learning age where several predictive applications that are life dependent are made by machines and robotic devices that relies on ensemble decision making algorithms. These have attracted many researchers and led to the development of an algorithm that is based on the integration of EKF, RBF networks and AdaBoost as an ensemble model to improve prediction accuracy. Firstly, EKF is used to optimize the slow training speed and improve the efficiency of the RBF network training parameters. Secondly, AdaBoost is applied to generate and combine RBFN-EKF weak predictors to form a strong predictor. Breast cancer survivability and diabetes datasets used were obtained from the UCI repository. Results are presented on the proposed model as applied to Breast cancer survivability and Diabetes diagnostic predictive problems. The model outputs an accuracy of 96% when EKF-RBFN is applied as a base classifier compare to 94% when Decision Stump is applied and AdaBoost as an ensemble technique in both examples. The output accuracy of ensemble AdaBoostM1-Random Forest and standalone Random Forest models is 97% respectively. The study has gone some way towards enhancing our knowledge and improving the prediction accuracy through the amalgamation of EKF, RBFN and AdaBoost algorithms as an ensemble model.

**Keywords:** AdaBoost, Breast Cancer, Diabetes Diagnosis, EKF, Ensemble, RBFN, Optimization, RMSE.

# **Introduction**

Ensemble models play crucial roles in many applications and related devices that are automated with the use of decision control mechanisms. Many ensemble algorithms are essentially iterative, and their results are inconsistent and not as accurate as it should be. Therefore, the need to develop enhanced predictive ensemble models are very important to the acceptability devices in the health and other industrial sectors that relies on them. However, the way algorithms are designed and trained plays a major role in machine learning prediction’s accuracy and reliability. In the past three decades the need for this have attracted many researchers that led to the development of wide range of approaches and variants of ensemble algorithms. However, the potential improvement in ensemble prediction through merging of the existing predictive models to improve classification accuracy has not been fully studied. In general, ensemble method combines several hypotheses (weak learners) to produce a strong classifier instead of the traditional standalone algorithms that are based on a single classifier. The selection and the diversity of the selected hypotheses also plays important role in prediction accuracy and reliability of ensemble models.

 One of the main objectives of ensemble machine learning algorithms as addressed in this research is to build and combine multiple weak learners on the same task to stabilize the prediction accuracy and achieve a better generalization result. For example, the accuracy prediction of breast cancer survival and diabetes diagnosis using data mining techniques based on historical records of patients can save lives by assisting doctors and policy makers in managerial decisions. It can also reduce the overhead cost of healthcare and other public service provisions*.* The proposed model is an extension of AdaBoost algorithm for forming committee of decision makers. The rationale behind this is that it takes the advantage of AdaBoost’s high bias RBFN’s (Radial Basis Function Network) noncomplex design. It also has good generalization, strong tolerance to input noise and EKF’s (Extended Kalman Filter) quicker convergence during iterations in addressing complex estimation problems.

 This paper therefore annexed this problem by proposing a concept that implements the process of integrating EKF, RBFN and AdaBoost algorithms as an ensemble model for binary classification tasks. A substantial additional output of this paper is the creation of working computerised EKF-RBFN-AdaBoost ensemble models. The models were evaluated and used as a computer assisted diagnosis device for early prediction of breast cancer and diabetic diagnostic diseases. The rest of the paper is arranged in the following format: In section 2 we provided an overview background of the problem. In section 3 we present an outline of algorithms that were integrated in the model proposed in this paper. Section 4 covers the results of our investigation and discussion of our findings. Finally, in section 5 we present conclusion of the model we proposed in this study and further work to be carried out in the future.

# **Background and Problem overview**

Ensemble algorithms are essentially iterative, their results are inconsistent and not as accurate as it should be. For example, the application of algorithms in early prediction of breast cancer and diabetes which are two common diseases that a lot of peoples requires algorithms with high prediction accuracy and reliability.

Breast cancer is one of the most common causes of cancer related death among women in the world in the past years. In the USA alone in 2015 an estimated 231,840 new cases of invasive breast cancer were diagnosed among women and 60,290 additional cases of in-situ breast cancer (Society American Cancer, 2015; Adegoke, Chen, & Banissi, 2017). Similarly, in the UK over 55,222 women were diagnosed with new cases of the disease in 2014 which amounted to 11, 433 deaths (UK, 2018) and the ailment reached 25.2% of women worldwide (Kwon & Lee, 2016). The disease is also a looming epidemic in the developing countries where advanced techniques for early detection and treatments are not readily available (Formenti, Arslan, & Love, 2012; Adegoke, Chen, & Banissi, 2017).

Similarly, “Diabetes is a chronic progressive disease that is characterized by elevated levels of blood glucose. Diabetes of all types can lead to complications in many parts of the body and can increase the overall risk of dying prematurely” (WHO, 2016). According to the British Heart Foundation “the increasing number of people suffering from the epidemic could trigger a 29% rise in the number of heart attacks and strokes linked to the condition by 2035” (BHF, 2018; ITV, 2018). Currently, about four million people in the UK have diabetes with condition accounting for 10% of all NHS spending (BBC, 2018).

## **2.1 Related Work**

Even though considerable research has been carried out in data mining using different ensemble techniques in predicting probable events based on historical datasets. One of the key challenges is the choice of the base classifier and appropriate loss function that goes with it. The goal of any ensemble algorithm is to minimize error rate to achieve accuracy and improve reliability. Despite the successful research efforts and application of ensemble methods (Adegoke, Chen, & Banissi, 2017), recent work shows that the problem with prediction accuracy, speedy and computational cost are still puzzling tasks. Therefore, the development of reliable ensemble models that can be applied for efficient medical diagnosis, incidents management and execution of automated technologies that are decision based and in some cases life dependent are highly essential. To address the issue of prediction accuracy/reliability and to extend the applications of ensemble algorithms, we propose a new model that bridges the potentials of RBFN, EKF and AdaBoost algorithms.

## **2.2 Breast Cancer Survivability Models**

Medically, breast cancer can be detected early during screening examinations through mammography or after a woman notices an unusual lump (Society American Cancer, 2015) in her breast. Owing to advancement in technology and availability of patient medical records, computer aided diagnosis cancer detection systems have been developed to detect and thus control the spread of the disease (Adegoke, Chen, & Banissi, 2017). However, such systems rely on pattern recognition algorithms that are used to process and analyse medical information of images obtained from mammograms for diagnostic and decision making (Weedon-Fekjær, Romundstad, & Vatten, 2014; Sapate & Talbar, 2016).

Different algorithms have also been proposed to extract relevant patterns from patients’ breast cancer datasets; for instance Yang et al (Yang, Lin, Chuang, & Chang, 2013) came up with a genetic algorithm that identify the relationship between genotypes that can lead to cancer cases using mathematical analysis. Similarly, McGinley *et al* (McGinley, et al., 2010; Adegoke, Chen, & Banissi, 2017) applied Spiking Neural Networks algorithm as a novel tumour classification method in classifying cancer tumours as either benign or malignant. In another approach (Pak, Kanan, & Alikhassi, 2015) proposed a breast cancer detection and classification in digital mammography based on Non-Subsampled Contourlet Transform (NSCT) and Super Resolution was proposed to improve the quality of digital mammography images. The authors then applied AdaBoost algorithm to determine the probability of a disease being a benign or malign cancer. Likewise, in breast mass cancer classification (Xie, Li, & Ma, 2015) the authors used computer-aided diagnosis (CAD) system for the processing and diagnosis of breast cancer. In their work, Adegoke *et al* proposed standalone and ensemble predictive models using AdaBoost as a technique and several base/standalone classifiers. The authors found that the topology and complexity of the algorithms does not necessarily improve the prediction performance of the models (Adegoke, Chen, & Banissi, 2017).

## **2.3 Diabetes Diagnostic Models**

In their study Alghamdi *et al* using SMOTE and ensemble techniques carried out experimental work using a number of algorithms to establish and compare their performances in predicting diabetes based data obtained from patients’ medical history (Alghamdi, et al., 2017). The model comprises of ensemble-based predictive method that uses 13 out of the 62 available classified attributes. The selected attributes for the model depends on clinical importance, multiple linear regression (MLR) and the Information Gain (IG). The authors reported an accuracy of 89% for G1/G2 and attributes and accuracy (AUC) of 0.922 for the ensemble method. Similarly, in (Zheng, et al., 2017), the authors proposed a framework that identifies type 2 diabetes using patient’s medical data. They utilized various classification models that extract features to predict identification of T2DM in datasets. According to the authors, the average results of the framework was 0.98 (UAC) compare with other algorithms at 0.71. To validate whether there is a connection between diabetes mellitus and glaucoma chronic diseases, in their work (Apreutesei, et al., 2018) applied a simulation technique constructed using artificial neural networks on clinical observations datasets*.* According to the authors the model was able to predict an accuracy of 95%. In another study (Barakat, Bradley, & Barakat, 2010) the authors proposed a multi-purpose model for the diagnosis and prediction of diabetes using support vector machines algorithm. The results of the model show a prediction accuracy of 94%, precision of 94%, and sensitivity of 93%.

# **3.0 RBFN, EKF and AdaBoost Algorithms**

Review on ensemble modelling reveals that there are substantial gaps in the literature where EKF algorithm could be used to train RBFN to form committee of EKF-RBFN model using AdaBoost. The performance of radial basis function network is based on how the network is trained and the training parameters obtained. Though, EKF have been used for modelling and calibration of dynamic systems such as model-based engine control architecture, ballistic and other space-based projects (Csank & Connolly, 2016) because of its performance even when noises are present. Despite the reliability and advanced application EKF and RBFN and the benefits they offered individually they have not been integrated together with AdaBoost as an ensemble model.

## **3.1 Radial Basis Function (RBF) Network**

RBF network is a special type of MLP (Multi-Layer Perceptron) artificial neural network for non-linear modelling (Nabney, 2002). The commonly used activation function for the network is radial basis function, however other functions such as Multiquadric or Thin-plate spline can similarly be applied. The output of RBF network is a linear combination of the radial basis functions of the inputs and neuron parameters that form part of the training process of the network. In most cases it uses radial basis function as an activation function and has only one hidden layer. However, it can be trained in many ways, unlike the MLP that are typically trained with backpropagation algorithms. The structure of a typical RBF network is shown in Figure 1. The output of the network can be expressed as in Equation 1

|  |  |  |
| --- | --- | --- |
|  | $$y\left(x\right) = \sum\_{j=1}^{M}w\_{j}ϕ\_{j}+ w\_{k} $$ |  (1) |
|  | $$ϕ\_{j}=ϕ\left(\left|\left|x- c\_{j}\right|\right|\right)$$ |  (2) |

where, $w\_{j}$ is the weight of $j^{th}$ centre,$ϕ\_{j}$ are the basis functions and $w\_{k}$ are the bias weights and $|\left| x- C\_{j}|\right|$ as expressed in Eq. 2 is the Gaussian activation function.

## **3.2 Extended Kalman Filter**

Theoretically, Kalman filter is a recurrence algorithm with a number of equation that can be used to estimate the state of a process that is based on series of measurements taken over time. The mean square error of the filter is minimised even when the measurements taken contains noises or missing data. The filter has been used in training neural network (Lima, Sanches, & Pedrino, 2017; Chernodub, 2014). The derivation and application of EKF are widely available in the (Ribeiro, 2004). It uses several measurements observed over time that contains noises and other inaccuracies. The filter consists of a number ensemble equation as illustrated in Figure 2. The filter also produces estimates of unknown variables that is more precise than those based on a single measurement. It minimizes the estimated covariance error in a Gaussian environment. When the conditions of Gaussian are not met it has been found that Kalman Filter still outperforms other class of linear unbiased filters (Merwe, Nelson, & Wan, 2004). Extended Kalman Filter on the other hand is the nonlinear form of the Kalman filter that linearizes the estimate of the current mean and covariance. Due to EKF’s simplicity and accuracy of predictions even in the presence of noise the algorithm has been widely accepted and used among scholars as a standard in the concept of GPS, nonlinear state estimation and other related complex nonlinear problems (Moreno & Pigazo, 2009).

## **3.3 AdaBoost as an ensemble technique**

AdaBoost an ensemble technique forms a strong classifier by combining the outputs of the weak classifiers (Adegoke V. F., Chen, Barikzai, & Banissi, 2017). It has many potential applications it has been successfully applied in many areas such as text classification, natural language processing, drug discovery and computational biology (Fan, Zheng, & Li, 2015) vision and object recognition (Viola & Jones, 2004; Lee, Han, & Ko, 2013), medical diagnosis (Abuhasel, Iliyasu, & Fatichah, 2015) and industrial chemical fault diagnosis (Karimi & Jazayeri-Rad, 2014). The key objective of AdaBoost as a meta-classifier is to improve the accuracy of the base (weak) classifiers by constructing and combining multiple instances of a weak classifiers (Schapire & Freund, 2014; Adegoke V. F., Chen, Barikzai, & Banissi, 2017) that produces a strong classifier that performs better than arbitrary guessing. Each instance of the classifier is trained on the same training dataset with different weights assigned to each instance based on its classification accuracy. The final classifier $H(x)$ as illustrated in Figure 3 is computed as a weighted majority of the weak hypothesis $h\_{t}$ by vote where each hypothesis is assigned a weight$ α\_{t}$ can be expressed as illustrated in equation 3 below.

|  |  |  |
| --- | --- | --- |
|  | $$H\left(x\right)=sign\left(\sum\_{t=1}^{T}α\_{t}h\_{t}\left(x\right) \right) $$ |  (3) |

During the training procedure there is a difference between the predicted values and the expected values. This difference over the committee of classifiers can be stated mathematically as in equation 4:

|  |  |  |
| --- | --- | --- |
|  | $$E\_{err}= \frac{1}{N}\sum\_{i=1}^{N}\left[y\_{i}\left(x\right)-h\_{i}\left(x\right)\right]^{2} $$ |  (4) |

## **3.4 Optimizing RBFN training parameters using EKF**

Kalman filter can be used to optimize weight matrix and error centre of RBFN as a least squares minimization problem (Simon, 2002). Therefore, in this session emphasis is on how EKF algorithm can be applied to minimize errors when used in training RBFN to improve their performance following a similar approach (Adegoke V. F., 2018). Assuming a non-linear finite dimension discrete time system we can represent the state and measurements as:

|  |  |  |
| --- | --- | --- |
|  | $$θ\_{k+1}=f\left(θ\_{k}\right)+ ω\_{k}$$ |  (5) |
|  | $$y\_{k}=h\left(θ\_{k}\right)+ v\_{k}$$ |  (6) |

where: the vector $θ\_{k}$ is the state of the system at time$ k$, $ω\_{k}$ is the process noise, $y\_{k}$ is the observation vector, $v\_{k}$ is the observation noise, $f\left(θ\_{k}\right) $and $h\left(θ\_{k}\right)$ are the non-linear vector functions of the state and process respectively (Adegoke V. F., 2018). If the dynamic models$f\left(θ\_{k}\right) $and $h\left(θ\_{k}\right)$ in equations 5 and 6 are assumed to be known. Then EKF can therefore be used as the standard technique of choice in approximating the maximum likelihood estimation of the state $θ\_{k}$ (Wan & Van Der Merwe, 2000). The state and the output white noises $ω\_{k}$ and $v\_{k}$ have zero-correlation with covariance matrix $Q\_{t}$ and $R\_{t}$ respectively and can therefore be modelled as:

|  |  |  |
| --- | --- | --- |
|  | $$Q=E\left[ω\_{k}ω\_{k}^{T}\right]$$ |  (7) |
|  | $$R=E\left[v\_{k}v\_{k}^{T}\right]$$ |  (8) |
|  | $$MSE=E\left[e\_{k}e\_{k}^{T}\right]= P\_{k}$$ |  (9) |

*where* $P\_{k}$ *is the error covariance matrix at time* $k.$

The solution that Kalman filter provide is to find an estimate for $\hat{θ}\_{n+1}$ from $θ\_{k+1}$ given$ y\_{j }(j=0,…,k)$. If the EKF model in Eq. 5 and Eq. 6 are further assumed to be sufficiently smooth, then we can expand them and approximate them around the estimate $θ\_{k}$ using first-order Taylor expansion series such that:

|  |  |  |
| --- | --- | --- |
|  | $f\left(θ\_{k}\right)=\left(\hat{θ}\_{k }\right)+F\_{k}\*\left(θ\_{k}-\hat{θ}\_{k }\right)$ + Higher orders |  (10) |
|  | $f\left(θ\_{k}\right)=\left(\hat{θ}\_{k }\right)+ H\_{k }^{T}\* \left(θ\_{k}-\hat{θ}\_{k }\right)$ + Higher orders |  (11) |

where

|  |  |  |
| --- | --- | --- |
|  | $$F\_{k}= \frac{∂f\left(θ\right)}{∂\left(θ\right)}|\_{θ=\hat{θ}\_{k }}$$ |   (12) |
|  | $$H\_{k }^{T}= \frac{∂h\left(θ\right)}{∂\left(θ\right)}|\_{θ=\hat{θ}\_{k }}$$ |   (13) |

Removing the higher order terms of the Taylor series and substitute Eq. 10 and Eq. 11 and substitute into Eq. 5 and Eq. 6 respectively, then Eq. 5 and Eq. 6 can be approximated as Eq. 14 and Eq. 15 respectively

|  |  |  |
| --- | --- | --- |
|  | $$θ\_{k+1}= F\_{k}θ\_{k}+ω\_{k}+ ∅\_{k}$$ |  (14) |
|  | $y\_{k}= H\_{k }^{T}+ v\_{k}+$ $φ\_{k}$ |  (15) |
|  |  |  |

# **4.0 Experimental Results and Discussion**

In this section we briefly describe the integration of RBFN, EKF and AdaBoost algorithms that were applied to enhance the prediction accuracy of the ensemble models we proposed in our study.

## **4.1 Improving RBFN-EKF Prediction models with AdaBoost**

In our simulation we fit the weak classifier RBFN to a version of the dataset described in the previous section. We used EKF to train the RBFN at each iteration. The training process comprises of several training points $(X\_{i}, Y\_{i})$ where $X\_{i}, $ $\in X$ and $Y\_{i} \in \{-1, +1\},$ on round m,$ where t = 1, . .. T$. Then we calculate the weighted misclassification rate of the learner and update the weighting measure used in the next i.e. *round t + 1*. During the training process AdaBoost called the base classifier *T* times, in this case 20 times. As AdaBoost trains RBF network at each round, RBFN layers are optimized using EKF to train and update the network training parameters, namely the: standard deviation ($σ$), mean ($μ$) and the weights ($w$). The architectural flowchart of the of our model is as illustrated in Figure 4 and the framework is as depict in Figure 5. As shown in Figure 5, it is possible to swap the dotted part (i.e. RBFN parameter optimization) of the framework with other optimization methods such as training the network with Decoupled Kalman filter or Particle Swarm Optimization (PSO).

## **4.2 Experimental Results and Analysis**

Some of the results of applying the proposed model, *EKF-RBFN-AdaBoost* as described on Breast Cancer survivability and Diabetes Diagnostic datasets are presented in this section. To evaluate the prediction accuracy and performance of the proposed models, we compare their results with standalone and ensemble models on the same datasets. In doing this, the following evaluation measures were used: Overall Accuracy Error Rate, True Positive, False Positive and F-Measure; Sensitivity and Precision. Tables 1 and 2 depicts the performance of the proposed model on cancer survivability and diabetic diagnostic datasets when the network was trained with EKF algorithm, compare with ensemble and standalone models. As illustrated Figure 7 the prediction accuracy of the model outperforms other ensemble models apart from AdaBoost-Random Forest. Figure 6 also shows the FPR, Precision and F-measure of the model are 0.03, 0.97 and 0.87 respectively. Figures 8 and 9 shows the performance of other standalone algorithms when tested on the same dataset. Figures 10 and 11 demonstrates the performance of the proposed model on diabetic dataset when the network EKF was used to train the network, compare with other ensemble models. It appears from Figure 9 that prediction accuracy of the model was 0.76 and outperform other models apart from AdaBoost-Random Forest which was also 0.76. As Figure 8 indicates, the TPR (True Positive Rate), FPR (False Positive Rate), Recall, Precision and F-measure of the model are 0.74, 0.34, 0.74,0.74 and 0.74 respectively. Figures 12 and 13 illustrate the performance of other standalone algorithms when tested on the diabetes dataset.

# **5.0 Conclusion and Further Work**

This paper demonstrates how to train RBF networks with EKF algorithm to optimize the RBFN training parameters and improve ensemble prediction accuracy. We used AdaBoost to generate committee of weak learners of EKF-RBFN that are combined to form the final prediction. A performance comparison was carried out using Breast Cancer Survivability and Diabetes Diagnostic datasets that were obtained from the UCI repository. The result shows a good prediction outcome and fast convergence rate compared with standard standalone models and other ensemble RBFN models trained with K-means algorithm or Support Vector Machine. The proposed model improves the generalization and prediction accuracy of ensemble models. It also minimizes overfitting of the data and improve rate of convergence during training compare to other models used during our study. In the near future, further research will be focused on the application of the proposed model on complex/imbalance datasets, effect of diversity and algorithmic settings on prediction accuracy.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithms/Measures** | **TPR** | **FPR** | **Recall/Sensitivity** | **Precision** | **F-Measure** | **Accuracy**  |
| **Predictive Models of Ensemble Classifiers** |
| **EKF-RBFN-AdaBoost** | *0.93* | *0.03* | *0.80* | *0.97* | *0.87* | *0.96* |
| **AdaBoostM1 with Decision stump** | 0.94 | 0.08 | 0.94 | 0.94 | 0.94 | 0.94 |
| **AdaBoostM1 with RBFN trained with K-Means** | 0.96 | 0.04 | 0.96 | 0.96 | 0.96 | 0.96 |
| **AdaBoostM1 with Random Forest** | **0.97** | **0.04** | **0.97** | **0.97** | **0.97** | **0.97** |
| **AdaBoostM1 with Support Vector Machine** | 0.97 | 0.04 | 0.96 | 0.96 | 0.96 | 0.96 |
| **Predictive Models of Standalone Classifiers** |
| **Random Forest** | **0.97** | **0.04** | **0.97** | **0.97** | **0.97** | **0.97** |
| **Support Vector machine** | 0.97 | 0.03 | 0.97 | 0.97 | 0.97 | 0.96 |
| **K-NN** | 0.96 | 0.06 | 0.96 | 0.96 | 0.96 | 0.96 |
| **ANN** | 0.96 | 0.04 | 0.96 | 0.96 | 0.96 | 0.96 |
| **Naïve Bayes** | 0.96 | 0.03 | 0.96 | 0.97 | 0.96 | 0.96 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

Table 1 Prediction comparison of Wisconsin Cancer Survivability dataset

|  |
| --- |
|  Figure 1 The Topology of a Radial Basis Function Network |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithms/Measures** | **TPR** |  | **FPR** | **Recall/ Sensitivity** | **Precision** | **F-Measure** | **Accuracy**  |
|  | **Predictive Models of Ensemble Classifiers** |
| **EKF-RBFN-AdaBoost** | *0.74* |  | *0.34* | *0.74* | *0.74* | *0.74* | *0.76* |
| **AdaBoostM1 with Decision stump** | 0.74 |  | 0.35 | 0.74 | 0.74 | 0.74 | 0.74 |
| **AdaBoostM1 with RBFN trained with K-Means** | 0.74 |  | 0.34 | 0.74 | 0.74 | 0.74 | 0.74 |
| **AdaBoostM1 with Random Forest** | **0.76** |  | **0.32** | **0.76** | **0.76** | **0.76** | **0.76** |
| **AdaBoostM1 with Support Vector Machine** | 0.65 |  | 0.65 | 0.65 | 0.42 | 0.51 | 0.65 |
|  | **Predictive Models of Standalone Classifiers** |
| **Random Forest** | **0.76** |  | **0.31** | **0.76** | **0.75** | **0.76** | **0.76** |
| **Support Vector machine** | 0.65 |  | 0.65 | 0.65 | 0.42 | 0.79 | 0.65 |
| **K-NN** | 0.65 |  | 0.65 | 0.65 | 0.42 | 0.51 | 0.65 |
| **ANN** | 0.75 |  | 0.31 | 0.75 | 0.75 | 0.75 | 0.75 |
| **Naïve Bayes** | 0.76 |  | 0.31 | 0.76 | 0.76 | 0.76 | 0.76 |

Table 2 Prediction Comparison on Diabetes Diagnostic dataset

|  |
| --- |
|  |

Figure 2 Basic equations and process of Kalman Filter as a sequential ensemble method

|  |
| --- |
| Related image |

Figure 3 An ensemble model showing committee of weak neural network predictors



Figure 4 The Architectural Flowchart of the proposed EKF-RBFN-AdaBoost Model



Figure 5 The framework of the proposed ensemble model based on training RBFN with EKF showing the exchangeable node as illustrated in Fig. 4

|  |  |
| --- | --- |
| Figure 6 TPR, FPR and Recall | Figure 7 Precision, F-Measure and Accuracy |
| Figure 8 TPR, FPR and Recall | Figure 9 Precision, F-Measure and Accuracy |

|  |  |
| --- | --- |
| Figure 10 TPR, FPR and Recall | Figure 11 Precision, F-Measure and Accuracy |
| Figure 12 TPR, FPR and Recall | Figure 13 Precision, F-Measure and Accuracy |

# References

Abuhasel, K., Iliyasu, A., & Fatichah, C. (2015). A Combined AdaBoost and NEWFM Technique for Medical Data Classification. *Information Science and Applications, 339*, 801-809.

Adegoke, V. F. (2018). *Research report.* London South Bank University, Computer Science informatics, School of Engineering. London, UK: LSBU.

Adegoke, V. F., Chen, D., Barikzai, S., & Banissi, E. (2017). Predictive Ensemble Modelling: Experimental Comparison of Boosting Implementation Methods. *Manchester, European Modelling Symposium (EMS).* UK: IEEE. doi:10.1109/EMS.2017.13

Adegoke, V., Chen, D., & Banissi, E. (2017). Prediction of breast cancer survivability using ensemble algorithms. *International Conference on Smart Systems and Technologies (SST).* Osijek, Croatia: IEEE.

Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PLoS One, 12*(7). doi:10.1371/journal.pone.0179805

Anil, K., Duin, R., & Mao, J. (2000). Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 22*(1), 4 - 37. doi:10.1109/34.824819

Apreutesei, N. A., FilipTircoveanu, Cantemir, A., Bogdanici, C., Lisa, C., Curteanu, S., & Chiseliţă, D. (2018). Predictions of ocular changes caused by diabetes in glaucoma patients. *Computer Methods and Programs in Biomedicine, 154*, 183-190.

Barakat, N., Bradley, A. P., & Barakat, M. N. (2010). Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. *Transactions on Information Technology in Biomedicine, 14*(4).

BBC, U. (2018). *Labour's Tom watson 'reversed' type-2 diabetes through diet and exercise.* London: British Broadcasting Corporation. Retrieved September 12, 2018, from http://www.bbc.co.uk/news/uk-politics-45495384

BHF. (2018). *CVD Statistics – BHF UK Factsheet.* BHF(British Heart Foundation). Retrieved August 23, 2018, from https://www.bhf.org.uk/-/media/files/research/heart-statistics/bhf-cvd-statistics---uk-factsheet.pdf

Chernodub, A. (2014). Training Neural Networks for classification using the Extended Kalman Filter: A comparative study. *Optical Memory and Neural Networks, 23*(2), 96–103.

Csank, J. T., & Connolly, J. W. (2016). *Model-Based Engine Control Architecture With an Extended Kalman Filter.* The American Institute of Aeronautics and Astronautics. San Diego, California: NASA STI Program. Retrieved 02 06, 2018, from https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20160002248.pdf

Fan, M., Zheng, B., & Li, L. (2015). A novel Multi-Agent Ada-Boost algorithm for predicting protein structural class with the information of protein secondary structure. *J Bioinform Comput Biol, 13*(5). doi:10.1142/S0219720015500225

Formenti, S., Arslan, A., & Love, S. (2012). Global Breast Cancer: The Lessons to Bring Home. *International Journal of Breast Cancer*.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and application to boosting. *Journal of computer and system sciences, 55*, 119-139. doi:http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.8918

Haykin, S. (1996). *Adaptive Filter Theory* (3rd ed.). Prentice-Hall,.

ITV. (2018). *Surge in heart attacks and strokes predicted as diabetes epidemic takes its toll.* London, UK: ITV Report. Retrieved August 06, 2018, from https://www.itv.com/news/2018-08-23/surge-in-heart-attacks-and-strokes-predicted-as-diabetes-epidemic-takes-its-toll/

Karimi, P., & Jazayeri-Rad, H. (2014). Comparing the Fault Diagnosis Performances of Single Neural Networks and Two Ensemble Neural Networks Based on the Boosting Methods. *Journal of Automation and Control, 2*(1), 21-32.

Kwon, S., & Lee, S. (2016). Recent Advances in Microwave Imaging for Breast Cancer Detection. *International Journal of Biomedical Imaging*. Retrieved from http://dx.doi.org/10.1155/2016/5054912

Lee, Y., Han, D., & Ko, H. (2013). Reinforced AdaBoost Learning for Object Detection with Local Pattern Representations. *The Scientific World Journal, 2013*, 14 pages. doi:http://dx.doi.org/10.1155/2013/153465

Lima, D., Sanches, R., & Pedrino, E. (2017). Neural Network Training Using Unscented and Extended Kalman Filter . *Robotics & Automation Engineering Journal, 1*(4).

McGinley, B., O'Halloran, M., Conceicao, R., F.Morgan, M.Glavin, & E.Jones. (2010). SPIKING NEURAL NETWORKS FOR BREAST CANCER CLASSIFICATION USING RADAR TARGET SIGNATURES. *Progress In Electromagnetics Research C, 17*, 74-94. doi:doi:10.2528/PIERC10100202

Merwe, R., Nelson, A., & Wan, E. (2004). *An Introduction to Kalman Filtering.* OGI School of Science & Engineering lecture.

Moreno, V. M., & Pigazo, A. (2009). *Kalman Filter:Recent Advances and Applications.* Vienna, Austria: I-Tech Education and Publishing KG.

Nabney, I. (2002). *NETLAB Algorithms for Pattern Recognition.* (M. Singh, Ed.) London: Springer.

Pak, F., Kanan, H., & Alikhassi, A. (2015). Breast cancer detection and classification in digitalmammography based on Non-SubsampledContourlet Transform (NSCT) and Super Resolution. *Co m p u t e r Me t h o d s a n d P r o g r a m s i n B i om e d i c i n e, 122*, 89-107.

Ribeiro, M. (2004). *Kalman and Extended Kalman Filters: Concept, Derivation and Properties.* CiteSeer.

Sapate, S., & Talbar, S. (2016). Mammograms, An Overview of Pectoral Muscle Extraction Algorithms Applied to Digital. *Studies in Computational Intelligence*. doi:DOI 10.1007/978-3-319-33793-7\_2

Schapire, R., & Freund, Y. (2014). *Boosting: Foundations and algorithms* (2nd ed.). MIT Press.

Simon, D. (2002). Training Radial Basis Neural Networks with the Extended Kalman Filter. *Neurocomputing, 48*, 455-457.

Society American Cancer. (2015). *American Cancer Society. Breast Cancer Facts & Figures 2015-2016.* (C. DeSantis, R. Siegel, & A. Jemal, Eds.) Retrieved May 05, 2017, from American Cancer Society: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2015-2016.pdf

UK, C. R. (2018). *Cancer Research UK*. Retrieved 20 08, 2018, from Cancer Resarch UK: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer

Viola, P., & Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision, 57*(2), 137-154.

Wan, E., & Van Der Merwe, R. (2000). The Unscented Kalman Filter for Nonlinear Estimation, Communications, and Control Symposium. *Adaptive Systems for Signal Processing*, 153-158. doi:10.1109/ASSPCC.2000.882463

Weedon-Fekjær, H., Romundstad, P., & Vatten, L. (2014). Modern mammography screening and breast cancermortality: population study. *MMJ, 348:g3701*. doi:doi: https://doi.org/10.1136/bmj.g3701

WHO. (2016). GLOBAL REPORT ON DIABETES- World Health Organization. Geneva 27, Switzerland: WHO Library Cataloguing-in-Publication Data. Retrieved from http://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257\_eng.pdf;jsessionid=DFE5616C3480A8F293D9970CC0FA4EF1?sequence=1

Xie, W., Li, Y., & Ma, Y. (2015). Breast mass classification in digital mammography based on extreme learning machine. Neurocomputing, 73(3), 930-941. doi:https://doi.org/10.1016/j.neucom.2015.08.048

Yang, C.-H., Lin, Y.-U., Chuang, L.-Y., & Chang, H.-W. (2013). Evaluation of Breast Cancer Susceptibility Using Improved Genetic Algorithms to Generate Genotype SNP Barcodes. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10(2), 361-371. doi:10.1109/TCBB.2013.27

Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., . . . Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International Journal of Medical Informatics, 97, Pages 120-127.