International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 11 (2019) pp. XXX-XXX
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Improving Prediction Accuracy of Breast Cancer Survivability and Diabetes Diagnosis via RBF Networks trained with EKF models

Vincent F Adegoke [1*], Daqing Chen[2], Ebad Banissi[3] and Sofia Barsikzai[4]
[1, 2, 3, 4] Computer Science and Informatics, School of Engineering,
London South Bank University London,
United Kingdom

*corresponding author, adegokev@lsbu.ac.uk; chenq@lsbu.ac.uk; banisse@lsbu.ac.uk; barikas@lsbu.ac.uk

Abstract: The continued reliance on machine learning algorithms and robotic devices in the medical and engineering practices has prompted the need for the accuracy prediction of such devices. It has attracted many researchers in recent years and has led to the development of various ensembles and standalone models to address prediction accuracy issues. This study was carried out to investigate the integration of EKF, RBF networks and AdaBoost as an ensemble model to improve prediction accuracy. In this study we proposed a model termed *EKF-RBFN-ADABOOST*. It uses EKF to enhance the slow training speed and to improve the effectiveness of the RBF network training parameters. AdaBoost was then applied as an ensemble meta-algorithm to generate and combine several RBFN-EKF weak classifiers to form a final strong predictor of the model. Breast cancer survivability, diabetes diagnostic, credit card payment defaults and staff absenteeism datasets used in the study were obtained from the UCI repository. The prediction accuracy of the proposed model was explored using various statistical analysis methods. During the study we also proposed and developed an ensemble logistic regression model using the breast cancer dataset. Results are presented on the proposed model *EKF-RBFN-ADABOOST*, as applied to breast cancer survivability, diabetes diagnostic, credit card payment defaults and staff absenteeism predictive problems. The model outputs an accuracy of 96% when EKF-RBFN was applied as a base classifier compared to 94% when Decision Stump was applied and AdaBoost as an ensemble technique in both cases. Also, a significant performance was observed for staff absenteeism at 96 % compared with credit card payment defaults that had a performance accuracy of 85%. The ensemble logistic model outputs an accuracy of 94% when we used 70% and 30% as training and testing datasets respectively compared with accuracy of 95% prediction when we used 60% of the data for training and 40% for testing respectively.

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

I. Introduction

Ensemble algorithms play crucial roles in many applications and related devices that are operated with the use of decision control mechanisms. Studies show that many of such algorithms are essentially iterative, and that their results are inconsistent and not as accurate as it should be. Therefore, the need to develop an improved predictive ensemble models are very significant to the acceptability of such devices in the health care and other industrial sectors that relies on them. In addressing this, many researchers have devoted attention to the problem. This has led to the development of a wide range of approaches and variants of ensemble algorithms. However, there are still some problems, such as the need to further improve their prediction accuracy and minimize overfitting problems. This paper is an extended version of the work that was originally presented in European Modelling Symposium on Mathematical Modelling and Computer Simulation (Adegoke, et al., 2018).

In general, ensemble algorithm combines several weak learners to produce a strong classifier instead of the traditional standalone algorithms that are based on a single classifier. Study shows that the choice and the diversity of the selected weak classifiers plays important role in prediction accuracy and reliability of the ensemble models. Recent study further shows that the potentials in ensemble prediction models through the merging of existing benchmark algorithms to improve prediction accuracy has not been fully considered.

One of the main objectives of ensemble machine learning algorithms as addressed in this research is to propose a new algorithm termed *EKF-RBFN-ADABOOST* that integrates EKF, RBFN and AdaBoost as an ensemble model for improved binary classification tasks. The proposed model builds and combines several weak learners on the same task to stabilize the prediction accuracy and to achieve a better generalization result. The rationale behind the proposed model is that it takes the advantage of AdaBoost's high prediction accuracy, RBFN's (Radial Basis Function Network) noncomplex design and EKF's (Extended Kalman Filter) quicker convergence during iterations when addressing complex estimation problems. Therefore, enabling the model to have good generalization, strong tolerance to input noise and missing data.

A substantial additional output of this paper is the creation of a working computerized ensemble EKF-RBFN-AdaBoost and an Ensemble Logistic Regression models. The models were evaluated and used as a computer assisted diagnosis device for early prediction of breast cancer, diabetic diagnostic diseases, staff absenteeism, and credit card payment defaults on datasets obtained from the UCI repository. The analysis of the simulation results of the study shows that the proposed algorithm *EKF-RBFN-ADABOOST* as a promising modelling technique. The result further shows that the model outperforms some of the standard ensemble and standalone classifiers. The accuracy prediction of breast cancer survival and diabetes diagnosis using data mining techniques based on historical records of patients using the proposed model can save lives by assisting doctors and policy makers in managerial decisions.

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 presented an outline of algorithms that were integrated into the model proposed in this paper, the EKF, RBFN, AdaBoost and the logistic regression models. Section 4 covers the experimental setup, results of our investigation and discussion of our findings. Finally, in section 5 we present the conclusion of the models we proposed in this study, and further work to be carried out in the future.

II Background and Problem overview

Review shows that ensemble techniques have become a popular method applied to solve classification and predictive problems in order to improve the quality and robustness of ensemble systems (Ghosh & Acharya, 2011; Kuncheva, et al., 2006), however not without challenges and problems. Despite the fact that ensemble algorithms are essentially iterative, study shows that their results are inconsistent and not as accurate as it should be in many areas.

For instance, breast cancer which is one of the most common causes of cancer related death amongst women in the world in the past years, requires the integration of predictive models with adequate and reliable results. 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, et al., 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 (Cancer Research, 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, et al., 2012; Adegoke, et al., 2017). Similarly, “Diabetes is a chronic progressive disease that is characterized by elevated levels of blood glucose. Research shows that 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). To address this it also requires the development of reliable and accurate predictive models. 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 the condition accounting for 10% of all NHS spending (BBC, 2018).

Therefore, the application of ensemble algorithms (which are non-invasive) in early prediction of breast cancer and diabetes which are two common diseases that affects a lot of peoples both in the developing and developed countries can no longer be overlooked. There is an urgent need to develop and integrate predictive models that can meets the required levels of predictive accuracy to control these diseases. Even though AdaBoost, EKF and RBFN have proved to be impressive algorithms in many devices and predictive applications. However, there are some situations where standalone networks might not be able to produce the required predictive results when handling complicated tasks. Such as imbalance datasets, and tasks where very high prediction accuracy are required such as cancer and diagnostic diabetes predictions as previously highlighted. Research shows that AdaBoost is susceptible to outliers (Changxin, et al., 2014; Kobetski & Sullivan, 2015) and in some cases overfitting (Jin & Zhang, 2007; Saravanakumar & Thangaraj, 2019). On the other hand, RBF networks could suffer from slow training speed and low efficiency (Gan, et al., 2012) if proper training algorithms are not applied in optimizing the training parameters as such can affect the predictive accuracy of the network.

A. Related Work

Even though considerable research has been carried out in data mining tasks using different ensemble techniques in predicting probable events based on historical datasets. One of the key challenges is the choice of the base classifier, the suitable loss function that goes with it and in some cases the appropriate algorithm to train the base classifier. Review shows that the goal of any ensemble algorithm is to minimize the error rate in order to achieve required accuracy and improved reliability. Irrespective of the successful research efforts and application of ensemble methods (Adegoke, et al., 2017), recent work shows that the problem with prediction accuracy, speed and computational costs are still puzzling problems (Huang, et al., 2017) that needs attention in order to take full advantage of the potentials of ensemble techniques. 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 medical devices are highly essential. Hence, 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 as an ensemble technique.

B. Breast Cancer Survivability Models

Recent research reveals that 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. However, owing to the recent advancement in technology and availability of patient medical records, computer aided diagnosis cancer detection applications have been developed to detect and consequently control the spread of the disease (Adegoke, et al., 2017). Recent research also shows that many of such applications rely on pattern recognition algorithms that are used to process and analyze medical information of images obtained from mammograms for diagnostic and decision making (Weedon-Fekjær, et al., 2014; Sapate & Talbar, 2016). Similarly, Yang et al (Yang, et al., 2013) proposed a genetic algorithm that detects the association of genotype frequencies of cancer cases and no cancer cases based on statistical analysis. The authors analyzed the possible breast cancer risks using odds-ratio and risk-ratio analysis. Likewise, McGinley et al (McGinley, et al., 2010) applied

Spiking Neural Networks algorithm as a novel tumor classification method in classifying tumors as either benign or malignant cancer. The performance of the technique was rated to outperform the existing Ultra-wideband (UWB) Radar imaging algorithm.

Equally, different algorithms have also been proposed to extract relevant patterns from patient's breast cancer datasets for instance Yang et al (Yang, et al., 2013) came up with a genetic algorithm that identifies the relationship between genotypes that can lead to cancer cases using mathematical analysis. Also, in their work Adegoke *et al* proposed standalone and ensemble predictive models using AdaBoost as a technique and several base classifiers (Adegoke, et al., 2017). The authors found that the topology and complexity of the algorithms does not necessarily improve the prediction and performance accuracy of the models. In another approach (Pak, et al., 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, et al., 2015) the authors used computer-aided diagnosis (CAD) system for the processing and diagnosis of breast cancer. In their predicting irritable bowel syndrome, a disease that is common among children Kau et al employs the use of a wrapper method to determine the optimum sample attributes (Kaur, et al., 2019). Then using an ensemble approach that comprises of five models and meta-algorithm to form the final classifier. According to the authors the model achieves an accuracy of 93.75%.

In another study using an automatic breast cancer detection technique that was based on hybrid features for pathological images, using a 3-output convolutional neural network that gives better segmentation results. The authors then applied a support vector machine with improved generalization and classified pathological image as benign or malignant based on the relief method for feature selection. According to the authors the method performs better when compared with existing techniques with a classification accuracy of 96.7% and 0.983 as the area under the curve (Bychkov, et al., 2018). In another approach an SVM-based ensemble learning algorithm was used to reduce the diagnosis variance and increase diagnosis accuracy of breast cancer diagnosis. In the study, 12 SVM models that were based on hybridized Weighted Area Under the Receiver Operating Characteristic Curve Ensemble were used in experimentation. According to the authors, the model reduces the variance by 97.89% and increases accuracy by 33.34% in comparison to the best single SVM model on the SEER dataset (Wang, et al., 2018).

C. Diabetes Diagnostic Models

In their study *Alghamdi et al* using SMOTE and ensemble techniques, the authors carried out experimental work by applying several algorithms to establish and compare their performances in predicting diabetes using data obtained from patients' medical history (Alghamdi, et al., 2017). The model comprises of ensemble-based predictive methods that uses 13 out of the 62 available features. The selected attributes were based on patient's clinical importance, multiple linear regression (MLR) and the Information Gain (IG). The authors reported an accuracy of 89% for G1/G2 attributes and accuracy (AUC) of 0.922 for the ensemble method. Similarly, in their work (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) compared with other algorithms at 0.71. To validate whether there is a connection between diabetes mellitus and glaucoma chronic diseases, in their study the authors (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 their MOSAIC project (Dagliati, et al., 2018), they used a data mining technique to derive predictive models of type 2 diabetes mellitus (T2DM) complications based on electronic health record data of patients. The model was based on patient's records: gender, age, time from diagnosis, BMI, glycated hemoglobin, hypertension, and smoking habit. They used Logistic Regression algorithm with a stepwise feature selection. The model was able to predict the onset of retinopathy, neuropathy, or nephropathy at different time scenarios, at 3, 5, and 7 years from the first visit of the patient at the Hospital Centre for Diabetes. The authors reported an accuracy of up to 84% of the model.

Even though reviews show that there is correlation between diabetes mellitus and glaucoma chronic diseases that affects people mainly over the age of 40. However, there is no validated evidence to support this. To validate whether there is a connection between the two diseases, in their work (Apreutesei, et al., 2018) the authors applied a simulation method constructed on artificial neural networks which was used in combination with clinical observations. They used a sample of 101 eye samples with an open angle glaucoma associated with the patients that had diabetes mellitus. According to the authors the model was able to predict an accuracy of 95%. Likewise, in addressing diabetes which been reported as a major cause of hospitalization and mortality in Taiwanese hospitals, Li et al (Li, et al., 2018) proposed a model that estimates of the risks of type 2 diabetes among patients. The authors used the Cox proportional hazard regression model to derive risk scores. According to the authors: "For the one-, three-, five-, and eight-year periods, the areas under the curve (AUC) for diabetes-related hospitalization in the validation set were 0.80, 077, 0.76, and 0.74, respectively with a corresponding value for in-hospital mortality in the validation set were 0.87, 080, 0.77, and 0.76." Similarly, in their study (Barakat, et al., 2010), the authors proposed a hybrid model for the diagnosis and prediction of diabetes using support vector machines algorithm. According to the authors the extracted rules using the model are reported to agree with the outcome of appropriate medical studies. The results of the model on a diabetes dataset indicate that model shows a prediction accuracy of 94%, a sensitivity of 93%, and a specificity of 94%.

In their study *Zu et al* applied decision tree, random forest and neural network on patient's dataset to predict diabetes mellitus. Due to the unbalanced nature and size of the dataset the authors used principal component analysis and minimum redundancy maximum relevance to reduce dimensionality of the dataset. According to the authors random forest produces the highest accuracy of 81% when all the attributes were used in simulating the data (Zou, et al., 2018). In a similar approach, a deep learning method was applied for the classification of diabetic and normal HRV signals (Swapna, et al., 2018). A short-term memory (LSTM), Convolutional Neural Network (CNN) and its combinations were used to extract complex temporal dynamic features from the heart rate variability data. The features were then passed into support vector machine's (SVM) for classification. According to the authors the technique gives a performance improvement of 0.03% and 0.06% in CNN and CNN-LSTM respectively compared with similar models without the integration of SVM algorithm.

III EKF, RBF Network, AdaBoost and Logistic Regression Algorithms

The performance of radial basis function network is based on how the network is trained and how the training parameters are obtained. Review shows that 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) with good performance even when noises are present. Equally RBFN have also been used in real world applications with good results compared with other algorithms. Despite the reliability, advanced applications of EKF and RBFN and the benefits of the algorithms offered individually, review shows that the algorithms have not been integrated together with another meta-algorithm such as AdaBoost to form an ensemble predictive model. In this section we a give brief property description of EKF, RBFN, AdaBoost, and ensemble Logistic techniques that the models we proposed in this study were based on.

A. Radial Basis Function Network

RBF network is a type of multi-layer perceptron artificial neural network for non-linear modelling. The commonly used activation function for the network is Radial Basis Function Network (RBFN), other common functions such as Multiquadric or Thin-plate spline can similarly be applied. Similarly, other kernel functions as depicted in Table 1 could also be used in training the network. Recent study shows that researchers have trained RBF networks by random selection of the centers from the data while others have used unsupervised methods such as the K-means algorithm (Qiao, et al., 2016) in selecting the network centers. In addition, others have also used supervised methods such as Particle Swarm Optimization (PSO) (Kelwade & Salankar, 2016; Wang, et al., 2015) and Gradient Descent (Malathi & Suresh, 2014; Soni, et al., 2015) algorithms to determine the parameters of the network. However, in this paper we used EKF to train RBF networks to optimize the network training parameters before applying AdaBoost as a technique to form ensemble of EKF-RBF networks as presented in next section. 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. The structure of a typical RBF network is as shown in Figure 1. The output of the network can also be expressed as in Equation 1.

$$y(x) = \sum_{j=1}^{M} w_j \phi_j + w_k \qquad (1)$$

$$\phi_j = \phi\left(\left|\left|x - c_j\right|\right|\right) \qquad (2)$$

where, w_j is the weight of j^{th} centre,ϕ_j are the basis functions and w_k are the bias weights and $||\, x - C_j||$ as expressed in Equation 2 as the Gaussian activation function.

Table 1 Common Radial Basis Kernel Functions

Basic Function (Abbreviation)	**Formula** $\boldsymbol{\emptyset(r) = \emptyset(\,||x - \mu||)/\sigma}$	**Smoothness**
Gaussian (GA)	$\boldsymbol{e^{-cr^2}}$	Infinite
Generalized Multiquadratic (GMQ)	$\boldsymbol{(c^2 + r^2)^{\beta}}$	Infinite
Inverse Multiquadratic (IMQ)	$\boldsymbol{1/\sqrt{c^2 + r^2}}$	Infinite
Inverse Quadratic (IQ)	$\boldsymbol{(c^2 + r^2)^{-1}}$	Infinite
Multiquadratic (MQ)	$\boldsymbol{\sqrt{(c^2 + r^2)}}$	Infinite
Hyperbolic Secant (sech)	$\mathrm{sech}(cr)$	Infinite
Cubic (CU)	r^3	Piecewise
Linear (LI)	r	Piecewise
Monomial (MN)	r^{2k-1}	Piecewise
Thin Plate Spline (TPS)	$r^2\log(\mathrm{r})$	Piecewise

B. Kalman Filter as a training algorithm

Hypothetically, Kalman Filter is a recurrence algorithm with several equations that can be used to estimate the state of a process that is based on series of measurements taken over a period of time. The filter (Kalman, 1960) is an optimal estimator algorithm that can deduce unknown values of interest from inaccurate and uncertain observations. Even though it was originally developed as a recursive solution to the discreet data linear filtering problem, it has been used to estimate linear system models with additive independent white noises. Theoretically, the filter uses several measurements observed over time that contains noises and other inaccuracies which it filters to predict the future behaviour of a system based on the system's past behavior, taking into consideration the environmental constraints of the system. The Extended Kalman Filter (EKF) on the other hand is the nonlinear version of the Kalman Filter which linearizes the estimate of the current mean and covariance. The algorithm has been considered as a standard in the theory of nonlinear state estimation, navigation systems and other related problems. The filter is able to produce estimates of unknown variables that is more precise than those based on a single measurement. It also minimizes the estimated covariance error in a Gaussian environment. The mean square error of the filter is minimized even when the measurements taken contains noises or missing data. The filter has been used in training neural network (Lima, et al., 2017; Chernodub, 2014). The process of calculating the ensemble weights can be considered as a discreet and sequential estimation problem. Therefore, EKF as a sequential estimator can be applied to optimize the weights and parameters of the RBFN models as described above. The filter consists of number ensemble equations as illustrated in Figure 2. EKF was used in this study due to the non-linear nature of RFFN.

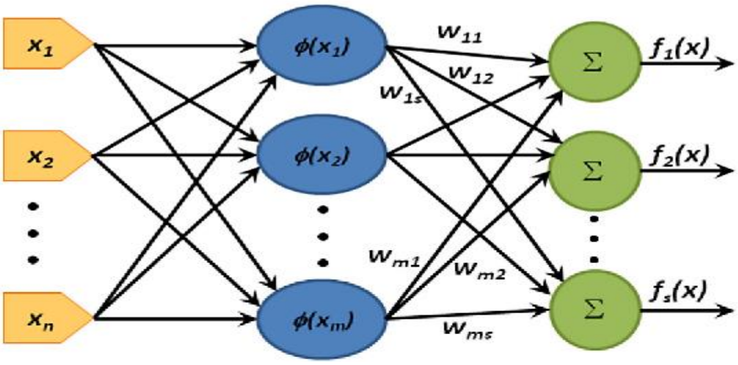
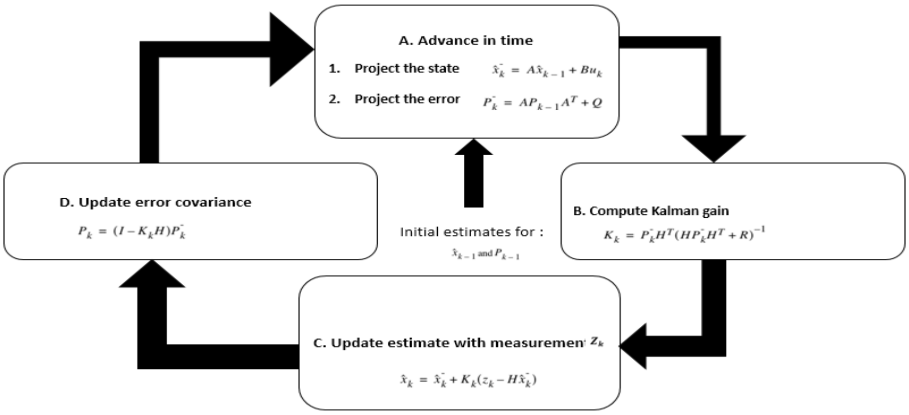


Figure 1 The Topology of a Radial Basis Function Network



A. Advance in time
1. Project the state $\hat{x}_k^- = A\hat{x}_{k-1} + Bu_k$
2. Project the error $P_k^- = AP_{k-1}A^T + Q$
B. Compute Kalman gain
$K_k = P_k^- H^T(HP_k^- H^T + R)^{-1}$
C. Update estimate with measuremen z_k
$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-)$
D. Update error covariance
$P_k = (I - K_k H)P_k^-$
Initial estimates for :
$\hat{x}_{k-1}$ and P_{k-1}

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

C. AdaBoost as an ensemble technique

AdaBoost is an ensemble technique that forms a strong classifier by combining the outputs of several weak classifiers on the same task. It has many potential applications and has been successfully applied in many areas such as text classification, natural language processing; drug discovery; computational biology (Fan, et al., 2015) vision and object recognition (Viola & Jones, 2004; Lee, et al., 2013), medical diagnosis (Abuhasel, et al., 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 classifiers by constructing and combining multiple instances of weak classifiers (Schapire & Freund, 2014; Adegoke, et al., 2017) and then producing a strong classifier that performs better than the arbitrary guessing.

The concept of AdaBoost is based on the idea that better algorithms can be created by combining multiple instances of a simple classifier. An ensemble model showing a committee of weak neural network predictors is as illustrated in Figure 4. The success of AdaBoost have been attributed to the algorithm's ability to reduce the training error and accelerate convergence after several iterations (Mukherjee, et al., 2013). Each instance of the base classifier is trained on the same training dataset with different weights assigned to each instance based on classification accuracy. AdaBoost's description here follows Schapire (Schapire, 2013) : assume we are given a number of labelled training examples such that $M = \{(x_1 , y_1), (x_2 , y_2), ., (x_n , y_n)\}$ where $x_i \in \mathcal{R}^M$ and the label $y_n \ \in \{-1, 1\}$. On each iteration $t = 1, \dots, T$, a distribution D_t is computed over the M training examples. A given weak learner is applied to find a weak hypothesis $h_t: \mathcal{R} \rightarrow \{-1, 1\}$. The aim of the weak learner is to find a weak hypothesis with low weighted error ε_t relative to D_t. The final classifier $H(x)$ is computed as a weighted majority of the weak hypothesis h_t by vote where each hypothesis is assigned a weight α_t. This is given in Equation 3:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \qquad (3)$$

The accuracy of the hypothesis is calculated as an error measure as depict Equation 4

$$\varepsilon_t \ = Pr_i{\sim}D_t[h_t(i) \neq \ y_i] \qquad (4)$$

The weight of the hypothesis is a linear combination of all the hypotheses of the participating as expressed in Equation 5

$$\alpha_t \ = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \qquad (5)$$

The distribution vector D_t of the weak classifiers is expressed as in Equation 6 where $\mathrm{Z_t}$ is a normalization factor such that the weights add up to 1 and makes $\mathrm{D_{t+1}}$ a normal distribution as illustrated in Equation 6.

$$D_{t+1}(i) = \frac{D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t} \qquad (6)$$

D. Some Theoretical Properties of AdaBoost

Some of the AdaBoost properties have been covered in several studies (Freud & Schapire, 2014). Therefore, we only highlight some of the properties that are relevant to our research in this section. Studies shows that it is possible to obtain an ensemble classifier with a lower exponential loss over training examples after each iteration such that the component classifier error is better than guess. This is illustrated in Equation 7 after expanding Equation 3.

$$\begin{aligned} &H_x\ (X) \\ &= \ \alpha_1 h(X_t) + \cdots \\ &+ \ \alpha_m h(X_m) \end{aligned} \qquad (7)$$

As depicted in Figure 3 the training classification error of the model must go down exponentially if indeed the weighted errors of the component classifiers are strictly better than guessing i.e. $\epsilon_k \leq 0.5$, the final hypothesis output of AdaBoost in Eq. 5.6 is bounded by Equation 8.

$$err\left(\hat{h}_m\right) \leq \prod_{k=1}^{m} 2\sqrt{\epsilon_k(1-\ \epsilon_k)} \qquad (4) \qquad (8)$$

Similarly, the weighted error of each new component classifier tends to increase as a function of the boosting iterations as shown in Equation 9. (5)

$$\epsilon_k \ = 0.5 - \frac{1}{2}(\sum_{i=1}^{n} \widehat{W_i}^{k-1} y_i h(X_i; \hat{\theta}_k)) \qquad (9)$$

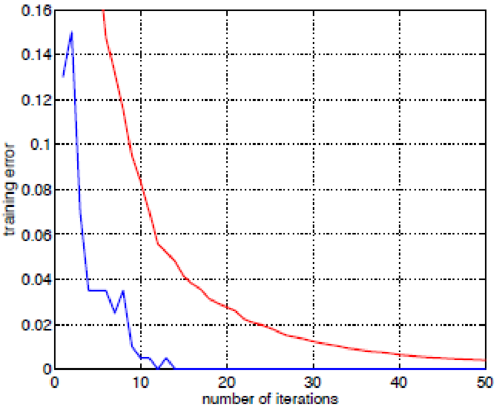


Figure 3 AdaBoost properties: training error

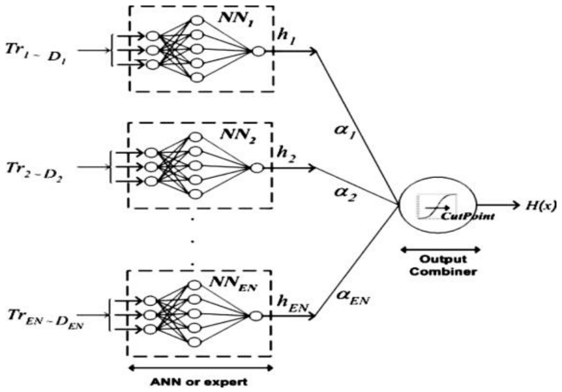


Figure 4 An ensemble model showing committee of weak neural network predictors

The expected test error i.e. the generalization error as presented in (Freund & Schapire, 1997; Freud & Schapire, 2014) has an upper bound with high probability and can be expressed as in Equation 10.

$$error_{true}(H) \le error_{train}(H) + \hat{O}\left(\sqrt{\frac{dT}{m}}\right) \qquad (10)$$

Where T is number of boosting rounds, d is the Vapnik-Chervonenkis (VC) dimension of weak learner that measures complexity of the classifier and m is the number of training examples. Review shows that AdaBoost has resistant to overfitting in practice. However, Equation 10 shows that if T is large AdaBoost will overfit. This means that the trained model can overfit the data and exaggerates variations in the data that can affect the generalization performance of the model. Study further shows that boosting increases the margin of classifier aggressively as it concentrates on the difficult examples during training rounds. Therefore, with large margin more weak learners and training rounds does not necessarily improve classification accuracy or increase the complexity of the final classifier. Despite this, boosting can still over fit if the boundary of separation is too small as weak learners can be too difficult to perform arbitrarily close to random guessing. According to Schapire et al (Schapire, et al., 1998) based on the concept of margin, given any threshold $\emptyset > 0$ of margin over data D, with a probability of $1 - \partial$, the generalization error of the ensemble $\in_D$ such that $\left(P_{x\sim D} f(x)\right) \neq H(x)$ is bounded by Equation 11.

$$\in_D \le P_{x\sim D}\big((fx)H(x) \le \emptyset +\big) + \hat{O}\left(\sqrt{\frac{d}{m\,\emptyset^2}} + \ln\frac{1}{\partial}\right) \qquad (11)$$

As we can see in Equation 11, it shows that as other variables are unchanged, then a larger margin over training data will lead to a smaller generalization error.

E. *Ensemble Logistic Regression Model*

The null hypothesis of a multiple logistic regression is that there is no connection between X variables and the predictable Y variables (McDonald, 2014). However, in multiple logistic regression there is a need to test a null hypothesis for each X variable to obtain the predictable Y variable, to show that adding X variable to the multiple logistic regression does not necessary improve the prediction accuracy of the equation. The main drive behind the use of multiple logistic regression is to determine the significant and the credible combination of the independent variables that best fit the dependent variable, such model can be expressed as in Equation 12.

$$E(Y_i \mid X_i) = \pi_i E(Y_i \mid X_i) = \frac{e^{(\beta_0 + \beta_1\beta_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})}}{1 + e^{(\beta_0 + \beta_1\beta_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})}} \qquad (12)$$

Where $\beta_0 \dots, \beta_i$ are the correlation coefficients and $X_{1i} \dots, X_{ki}$ are the variables and Y_i is like hood prediction for variables $X_{1i} \dots, X_{ki}$. Review shows that there are several methods that allows one to specify how the independent variables are chosen to form multiple regression models (McDonald, 2014; Mangiafico, 2015). Among the common techniques are the forward selection, the backward selection and the stepwise. In forward selection a single predictor that best fits the data is added to the equation, this is followed by adding other predictors that contributes significantly to the performance of the regression model one at a time. On other hand, in backward elimination all independent variables are added into the regression equation, then each variable is examined and removed one at a time if they do not contribute significantly to the regression equation. However, the stepwise regression is a mixture of the forward and backward selection methods that involves adding and removing variables to the model's equation. During our study we used a cancer dataset to build an ensemble logistic regression model for the prediction of cancer survivability. The plots correlations among the features of the dataset and the ordered variables with the highest correlation closest to the diagonal are as shown in Figure 5 and Figure 6 respectively. The statistical measures of the dataset using different features as a predictor is as shown in Table 2. To identify the prognostic factors and to develop an ensemble logistic regression model with multivariate features, we applied the $coef$ function in R programming package to extract the model's coefficients from the object returned by the modelling function. Some of the statistical properties of the model that were used in forming the ensemble logistic regression model displayed in equation 13 is as illustrated in Table 2.

$$ln(Y) = -10.10394 + 0.53501 * X1 - 0.00628 * X2 + 0.32271 * X3 + 0.33064 * X4 + 0.09663 * X5 + 0.38303 * X6 + 0.44719 * X7 + 0.21303 * X8 + 0.53484 * X9 \qquad (13)$$

Table 2 Statistical measures of Cancer dataset using different variables in predicting cancer prognosis

Data Features	**Accuracy**	**RSME**	**KAPPA**	**TP**	**FP**	**Precision**	**Recall**	**F-Measure**	**Features**
Clump thickness	85	0.324	0.651	0.855	0.260	0.874	0.989	0.899	X1
Uni Cell Size	92	0.240	0.823	0.919	0.076	0.923	0.919	0.920	X2

Uni Cell Shape	92	0.234	0.826	0.920	0.076	0.923	0.919	0.919	X3
Single epithelia	85	0.326	0.673	0.859	0.222	0.863	0.859	0.854	X4
Epithelial Cell Size	90	0.290	0.786	9.900	0.096	0.905	0.900	0.901	X5
Bare Nuclei	90	0.269	0.798	0.908	0.106	0.908	0.908	0.908	X6
Bland Chromatin	90	0.270	0.800	0.908	0.106	0.908	0.908	0.908	X7
Normal Nucleoli	89	0.302	0.769	0.898	0.146	0.897	0.898	0.896	X8
Mitoses	79	0.406	0.473	0.788	0.369	0.810	0.788	0.765	X9

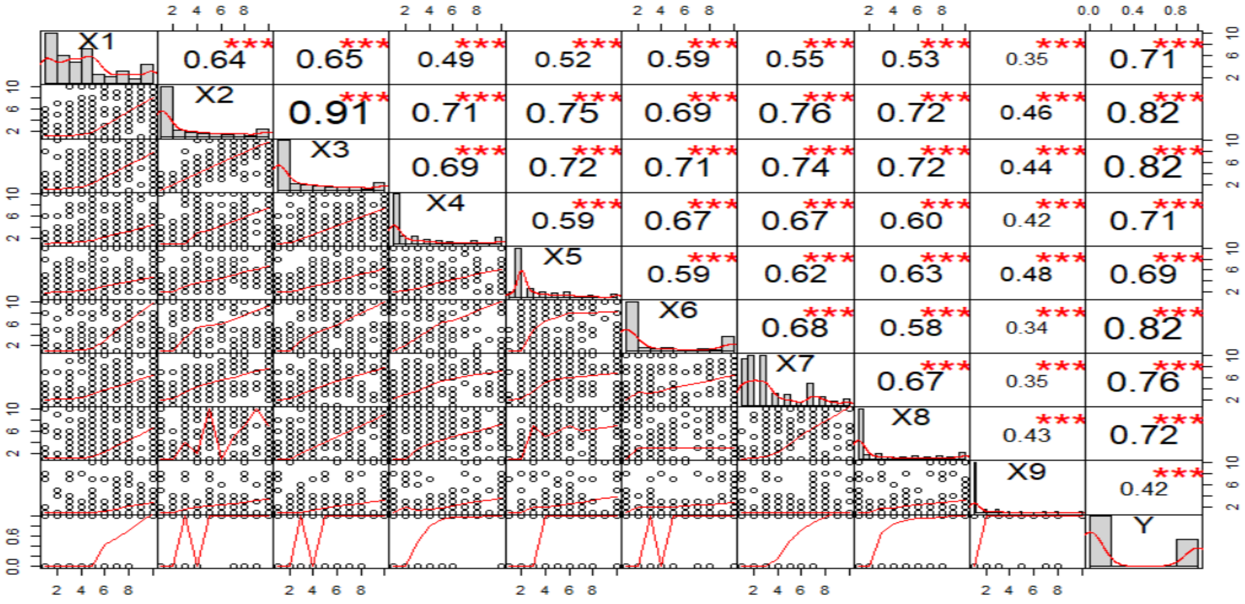


Figure 5 Correlations among the features of the Breast cancer dataset

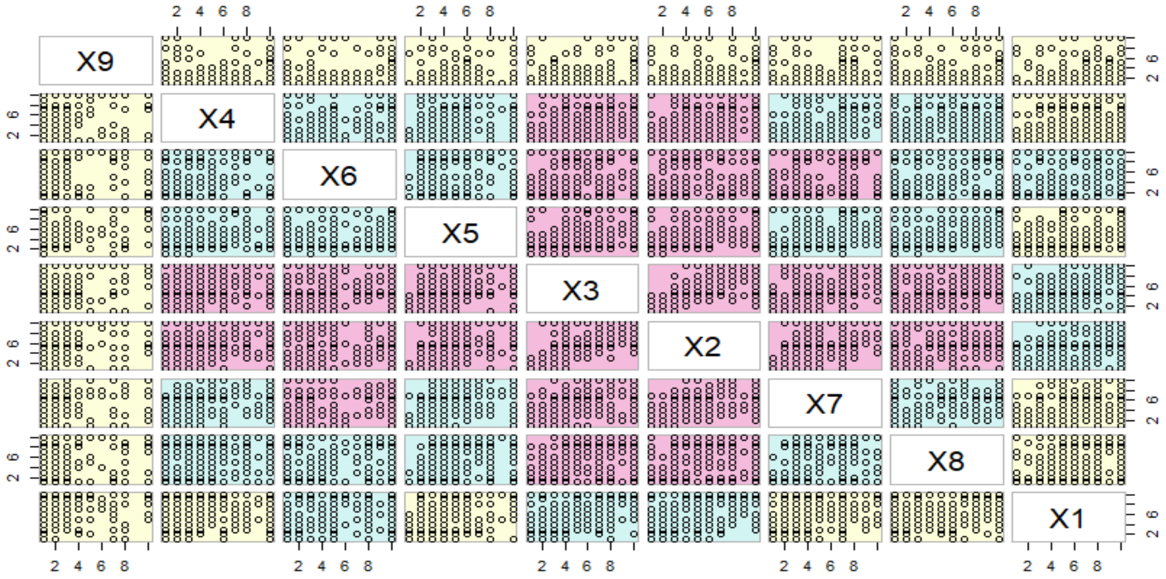


Figure 6 Ordered variables of cancer dataset with the highest correlation closest to the diagonal

Table 3 Statistical properties of the regression model as illustrated in Equation 1

	Estimate	**Std. Error**	**Z value**	**Significant**
Intercept	-10.10394	1.17488	-8.600	0
X1	0.53501	0.14202	3.767	0
X2	-0.00628	0.20908	-0.030	1

X3	0.32271	0.23060	1.399	1
X4	20.33064	0.12345	2.678	0.001
X5	0.09663	0.15659	0.617	1
X6	0.38303	0.09384	4.082	0
X7	0.44719	0.17138	2.609	0.001
X8	0.21303	0.11287	1.887	0.1
X9	0.53484	0.32877	1.627	1

```
# R-Code : Multinomial Regression for Cancer
Survivability Dataset
#Import data
setwd("H:/Res5/Datasets")
# Read CSV into R
cancerData <- read.csv(file="cancerDBlogit.csv",
header=TRUE, sep=",")
colnames(cancerData) <- c("X1", "X2", "X3", "X4",
"X5", "X6", "X7", "X8", "X9", "Y")
#head(cancerData)
#str(cancerData)
# Prepare Training and Test Data
set.seed(100)
#Training data 70%
trainingRows <- sample(1:nrow(cancerData), 0.7 *
nrow(cancerData))
trainingData <- cancerData[trainingRows, ]
#Test data 30%
testData <- cancerData[-trainingRows, ]
#Build Multinomial Model
library(nnet)
#multinom Model
multinomCancerModel <- multinom(Y ~ .,
data=trainingData)
# model the summary
summary(multinomCancerModel)
#Predict on Test Data
predicted_scores <- predict (multinomModel, testData,
"probs")
#Prediction on new data
predicted_class <- predict (multinomModel, testData)
#Confusion Matrix
table (predicted_class, testData$Y)
#Get the Misclassification Error as a percentage
MissClassError <-(mean(as.character(predicted_class)!=
as.character(testData$Y))) * 100
#Round the output to 2 decimal place and concatenate the
output with %
MissClassError <- paste (round (MissClassError, 2),
sep='', '%')
```

Figure 7 R-code: Training and testing the Ensemble Logistic Model

A snippet code of the R code used in training and testing the ensemble logistic model based on Cancer dataset is as illustrated in Figure 7. We found that the misclassification errors of the model are influenced by the percentage of dataset used in training and percentage used testing the model. Table 4 illustrates classification errors and the corresponding percentages of training and testing data of the model. This is graphically illustrated in Figure 8, while Table 5 shows a typical confusion matrix output of the model on based 70% of training data and 30% of testing data.

A. *Optimization of BFN training parameters with EKF*

As illustrated in the previous section the optimization of the ensemble weights is a type of discrete data filtering problem. Therefore, it is possible to use EKF to optimize the weight matrix in RBFN problems. Likewise, the training error of ensemble model can be treated as a least squares' minimization problem. The derivation and application of Kalman Filter as a sequential ensemble method are widely available in literature (Ribeiro, 2004). Review shows that only a few studies have examined the applications of EKF in training Neural Network (Haykin, 2008; Simon, 2002). Despite this, none of such studies have integrated such a solution with AdaBoost in generating ensemble of RBF network classifiers. In this session emphasis is laid on how EKF can be applied to optimize the training parameters of RBFN to improve their prediction performance. Assuming a non-linear finite dimension discrete time system we can represent the state and measurements as in Equations 14 and 15.

Table 4 Data Training size and classification error

Classification error	Training data %	Testing Data %
3.41	40	60
3.72	45	55
4.09	50	50
4.55	55	45
4.74	60	40
4.42	65	35
6.34	70	30
6.43	75	25
6.57	80	20
4.85	85	15
4.35	90	10

Table 5 Model Confusion Matrix

class	0	1
0	126	6
1	7	55

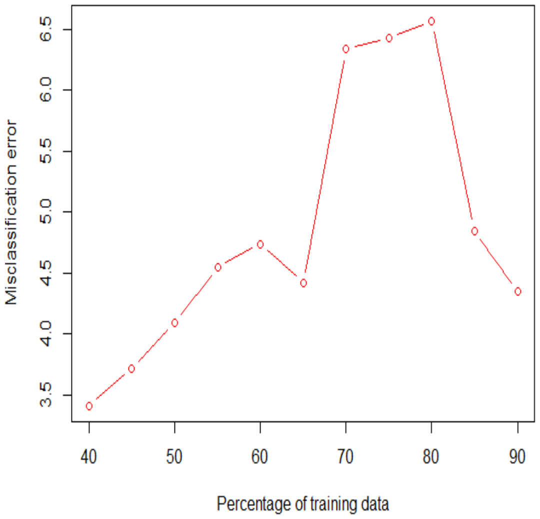


Figure 8 Percentage of training data vs classification error

$$\theta_{k+1} = f(\theta_k) + \omega_k \tag{14}$$

$$y_k = h(\theta_k) + v_k \tag{15}$$

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 and $f(\theta_k)$ and $h(\theta_k)$ are the non-linear vector functions of the state and process respectively. If the dynamic state $f(\theta_k)$ and process $h(\theta_k)$ in Equations 14 and 15 are assumed to be known, then EKF can be used as the standard method of choice to approximate maximum likelihood estimation of the state θ_k (Wan & Merwe, 2000). Consequently, applying similar approach as in (Puskorious & Feldkamp, 1994; Simon, 2002), we can view the optimization of RBFN with weight W and the prototype v_j as a weighed least-square minimization problem. The error vector can therefore be viewed as the difference between the RBFN outputs and the expected target values. The optimization problem of RBFN can therefore be represented using Extended Kalman Filter algorithm by letting the output of the weight W and the elements of the prototype v_j represent the state of a nonlinear system and the output of the RBFN network respectively. 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 be modelled as in Equation 16.

$$Q = E[\omega_k {\omega_k}^T] \tag{16}$$

$$R = E[v_k {v_k}^T] \tag{17}$$

$$MSE = E[e_k {e_k}^T] = P_k \tag{18}$$

where, P_k *is* the error matrix at time k.

Afterward, EKF aim to provide is to find an estimate for $\hat{\theta}_{n+1}$ from θ_{k+1} given y_j $(j = 0, \dots, k)$. If the EKF model in Equation 14 and Equation 15 are further assumed to be sufficiently smooth, then we can expand the equations and approximate them around the estimate θ_k using first-order Taylor expansion series such that:

$$f(\theta_k) = (\hat{\theta}_k) + F_k * (\theta_k - \hat{\theta}_k) \quad + \quad \text{Higher orders} \tag{19}$$

$$f(\theta_k) = (\hat{\theta}_k) + H_k^T * (\theta_k - \hat{\theta}_k) \quad + \quad \text{Higher orders} \tag{20}$$

where,

$$F_k = \frac{\partial f(\theta)}{\partial(\theta)}|_{\theta=\hat{\theta}_k} \tag{21}$$

$$H_k^T = \frac{\partial h(\theta)}{\partial(\theta)}|_{\theta=\hat{\theta}_k} \tag{22}$$

If drop the higher order terms of the Taylor series and substitute Equation 19 and Equation 20 into Equation 14 and Equation 15 respectively, then Equation 14 and Equation 15 can be approximated as Equation 23 and Equation 24 respectively.

$$\theta_{k+1} = F_k\theta_k + \omega_k + \emptyset_k \tag{23}$$

$$y_k = H_k^T + v_k + \varphi_k \tag{24}$$

Therefore, the estimated value $\hat{\theta}_n$ can be obtained using recursion as in (Simon, 2002) such that:

$$\hat{\theta}_k = f(\hat{\theta}_{k-1} + K_k[y_k - h(\hat{\theta}_{k-1})]) \tag{25}$$

$$K_k = P_k K_k (R + H_k^T P_k H_k)^{-1} \tag{26}$$

$$P_{k+1} = F_k(P_k - K_k H_k^T P_k)F_k^T + Q \tag{27}$$

where, K_k is the Kaman Gain, P_k is the covariance matrix of the estimation error, θ_{k+1} is state estimation, Q is the tuning parameter for ω_k (a covariance matrix), and R is the tuning parameter for v_k (which is also a covariance matrix).

IV Experimental Setup 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 this study. To evaluate the performance of the proposed model and to compare it with existing standalone and ensemble algorithms, some experimental case studies, and simulations were carried out based on benchmark datasets that were obtained from the UCI repository. The datasets are Wisconsin breast cancer survivability, diabetes diagnostic, staff absenteeism and credit card payment defaults. These case studies were performed using AdaBoost as an ensemble technique. We applied decision stump, K-means, random forest, support vector machine, ANN and Naïve Bayes as standalone algorithms. We also carried out experimental simulation on the cancer prognosis dataset using the ensemble logistic regression model described in the previous section.

A. Enhancement of RBFN-EKF Predictors

In the study we fitted the enhanced RBFN weak classifiers on the datasets as described in the previous section. EKF was applied in training 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 t, $where\ t = 1, \ldots T$. Then we calculated the weighted misclassification rate of the learner and update the weighting measure used in the next *round t + 1*. During the training process, AdaBoost called the base classifier T times, in our 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) using N different RBFN functions to generate different RBFN weak classifiers. The output of the model is the sum of the outputs of the several weak predictors trained by AdaBoost. The architectural flowchart of the model is as illustrated in Figure 9 and the framework is as depicted in Figure 10. As shown in Figure 10, it is possible to switch the dotted section (i.e. RBFN parameter optimization) of the framework with other parameter optimization algorithms such as Decoupled Kalman Filter, Particle Swarm Optimization (PSO) or with other training algorithms.

B. Experimental Results and Analysis

Some of the results of applying the proposed model, EKF-RBFN-AdaBoost are presented in this section. The following evaluation measures were used namely: Prediction Accuracy Error Rate, True Positive, False Positive and F-Measure; Sensitivity and Precision. Tables 6, 7, 8 and 9 depicts the performance of the proposed model described on breast cancer survivability, diabetes diagnostic, staff absenteeism and clients credit card payment default datasets compare with benchmark ensemble and standalone models. As can be seen in Table 6 the prediction accuracy of the proposed model on Cancer dataset is 96% compare with performance accuracy of 97% when Random Forest was use as base classifier with AdaBoost as prediction accuracy of 97% when random forest was used as a standalone algorithm. Likewise, in Table 7 the prediction accuracy of the proposed model on diabetics' dataset is 76%, as can be seen in the table this is the same prediction accuracy as Random Forest and ensemble AdaBoost + Random forest, however, the proposed model outperforms other models. Similarly, Table 8 illustrates the prediction accuracy of the proposed model compared with other methods on workers' absenteeism dataset. It shows that the performance accuracy of the model and that of the ANN are both 96%. The proposed model outperforms other models apart from Random Forest and AdaBoost + Random Forest which both have a prediction accuracy of 98%. Correspondingly, Table 9 shows the performance of the proposed algorithm based on credit card payment defaults with prediction accuracy of 85%. As can be seen in the table the model outperforms other predictive algorithms and techniques used in the study. The predictive performance of both Random Forest and ensemble Random Forest are 78% respectively. The performance of the proposed model on diabetes dataset compare with other models are as illustrated graphically in Figures 11, 12, 13, and 14. Likewise, the performance of the model on cancer dataset compare with other algorithms are as illustrated in Figures 15, 16, 17 and 18. Figures 19, 20, 21 and 22 also illustrates the performance of the proposed model on Absenteeism dataset compare with other machine learning methods. Similarly, Figures 23, 24, 25 and 26 demonstrates the performance of the proposed model on credit card payment default dataset compare with other machine learning methods. The performance of the proposed model on diabetes dataset compare with other models are as illustrated graphically in Figures 11, 12, 13, and 14. Likewise, the performance of the model on cancer dataset compare with other algorithms are as illustrated in Figures 15, 16, 17 and 18. Figures 19, 20, 21 and 24, 25 and 26 demonstrates the performance of the proposed model on credit card payment default dataset compare with other machine learning methods.

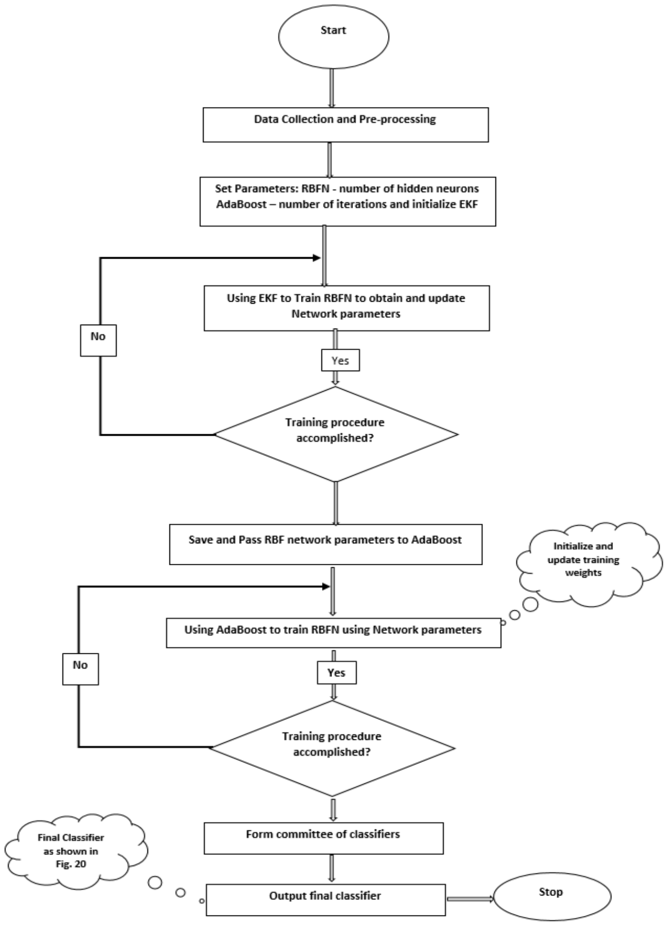

Start
Data Collection and Pre-processing
Set Parameters: RBFN - number of hidden neurons AdaBoost – number of iterations and initialize EKF
Using EKF to Train RBFN to obtain and update Network parameters
No
Yes
Training procedure accomplished?
Save and Pass RBF network parameters to AdaBoost
Initialize and update training weights
Using AdaBoost to train RBFN using Network parameters
No
Yes
Training procedure accomplished?
Form committee of classifiers
Final Classifier as shown in Fig. 20
Output final classifier
Stop

Figure 9 The Architectural flowchart of the proposed EKF-RBFN-AdaBoost Model

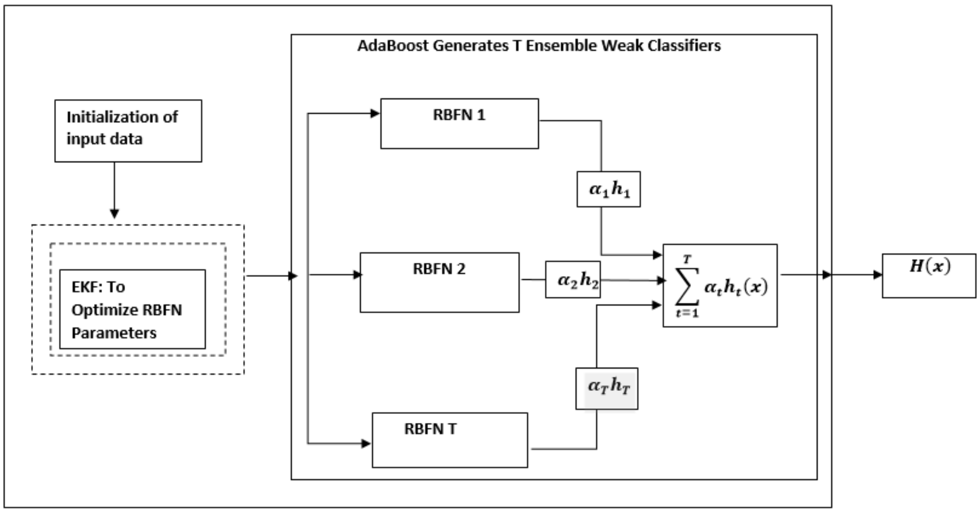

AdaBoost Generates T Ensemble Weak Classifiers
Initialization of input data
EKF: To Optimize RBFN Parameters
RBFN 1
RBFN 2
RBFN T
$\alpha_1 h_1$
$\alpha_2 h_2$
$\alpha_T h_T$
$\sum_{t=1}^{T} \alpha_t h_t(x)$
$H(x)$

Figure 10 The framework of the proposed ensemble model based on training RBFN with EKF showing the exchangeable node with dotted lines (that be integrated with other training algorithms such as PSO)

Table 6 Prediction comparison of Wisconsin Cancer Survivability dataset

Algorithms/Measures	**TPR**	**FPR**	**Recall**	**Precision**	**F-Measure**	**Accuracy**
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
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 7 Prediction Comparison on Diabetes Diagnostic dataset

Algorithms/Measures	**TPR**	**FPR**	**Recall**	**Precision**	**F-Measure**	**Accuracy**
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
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 8 Performance Comparison Using Workers Absenteeism

Algorithms/Measures	TPR	FPR	Recall	Precision	F-Measure	Accuracy
Predictive Models based on Ensemble Classifiers						
EKF-RBFN-AdaBoost	**0.95**	**0.85**	**0.95**	**0.94**	**0.95**	**96**
AdaBoostM1 + Decision stump	0.94	0.81	0.94	0.94	0.91	94
AdaBoostM1 + K-Means	0.94	0.52	0.94	0.93	0.93	94
AdaBoostM1 + with Random Forest	0.98	0.31	0.98	0.98	0.98	98
AdaBoostM1 + Support Vector Machine	0.91	0.72	0.91	0.90	0.90	92
Predictive Models Based Standalone Classifiers						
Random Forest	0.98	0.28	0.98	0.98	0.98	98
K-NN	0.98	0.52	0.94	0.93	0.93	94
Support Vector machine	0.92	0.91	0.92	0.88	0.89	92
ANN	0.97	0.34	0.97	0.96	0.97	96
Naïve Bayes	0.93	0.52	0.93	0.92	0.93	93

Table 9 Performance Comparison Using Clients Credit Card Defaults

Algorithms/Measures	TPR	FPR	Recall	Precision	F-Measure		Accuracy
Predictive Models based on Ensemble Classifiers							
EKF-RBFN-AdaBoost	**0.80**	**0.85**	**0.82**	**0.84**	**0.88**		**85**
AdaBoostM1 with Decision stump	0.80	0.59	0.81	0.78	0.78		81
AdaBoostM1 with RBFN trained with K-Means	0.73	0.55	0.73	0.73	0.73		73
AdaBoostM1 with Random Forest	0.79	0.77	0.79	0.73	0.70		78
AdaBoostM1 with Support Vector Machine	0.78	0.54	0.78	0.76	0.76		78
Predictive Models based on Standalone Classifiers							
Random Forest	0.78	0.76	0.79	0.75	0.70		78
Support Vector machine	0.78	0.54	0.78	0.76	0.74		78
K-NN	0.73	0.55	0.73	0.73	0.73		73
ANN	-	-	-	-	-		-
Naïve Bayes	0.53	0.43	0.53	0.70	0.57		53

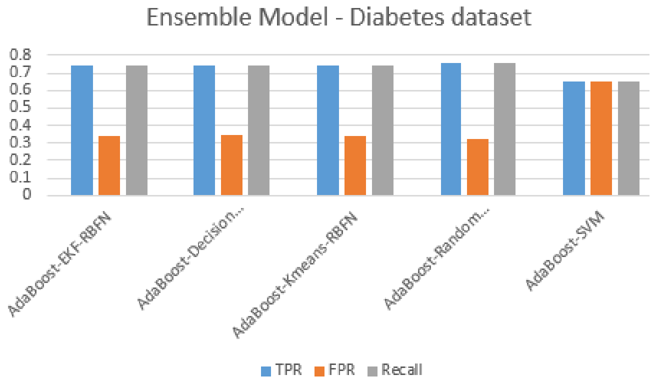


Figure 11 TPR, FPR and Recall

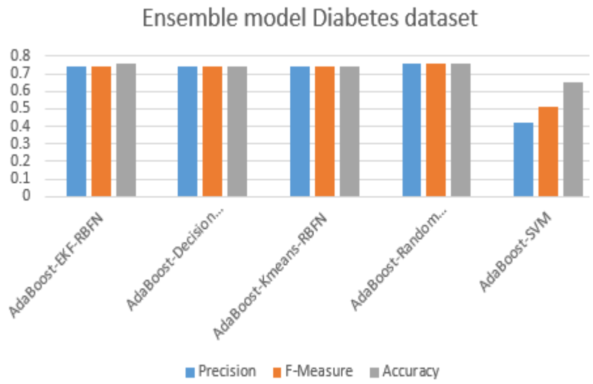


Figure 12 Precision, F-Measure and Accuracy

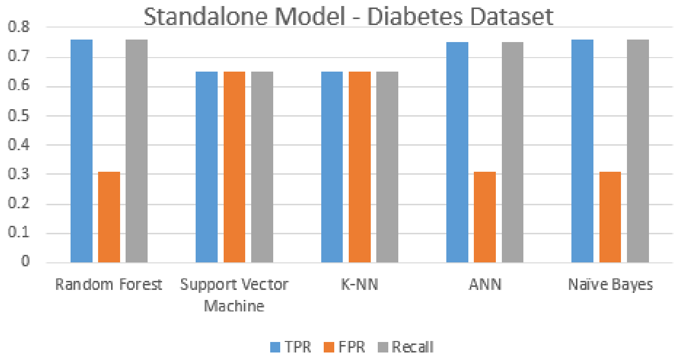


Figure 13 TPR, FPR and Recall

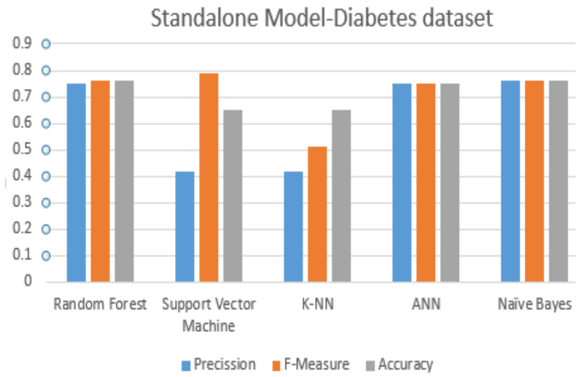


Figure 14 Precision, F-Measure and Accuracy

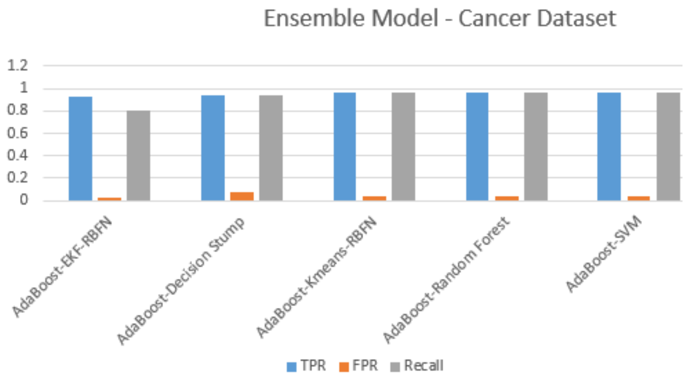


Figure 15 TPR, FPR and Recall

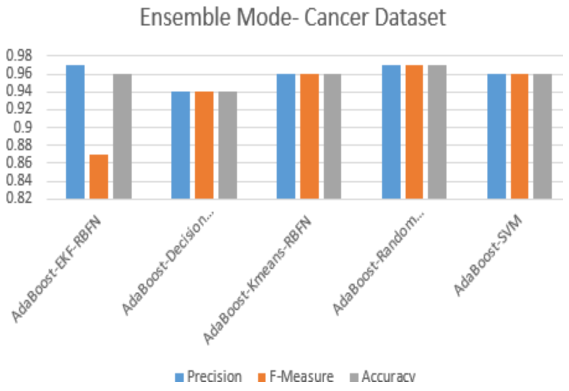


Figure 16 Precision, F-Measure and Accuracy

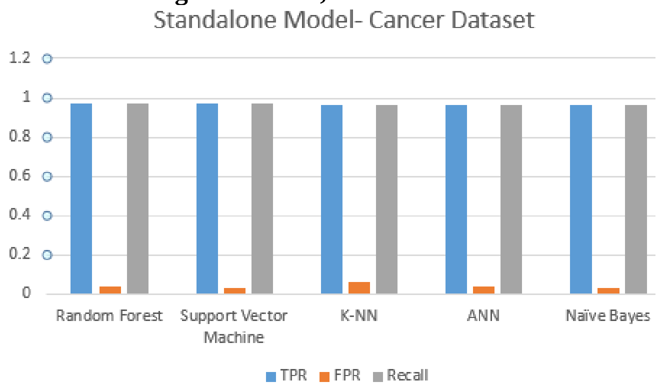


Figure 17 TPR, FPR and Recall

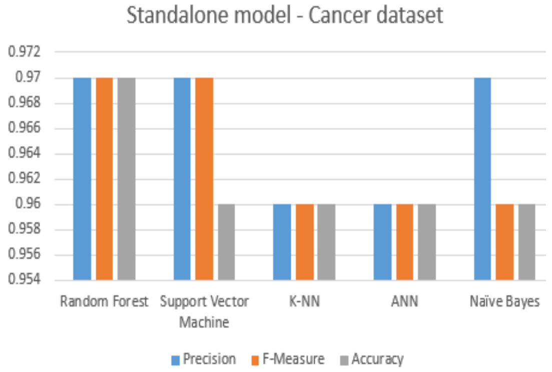


Figure 18 Precision, F-Measure and Accuracy

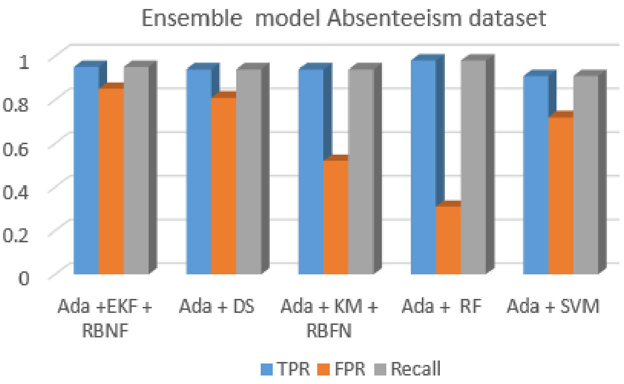


Figure 19 TPR, FPR and Recall

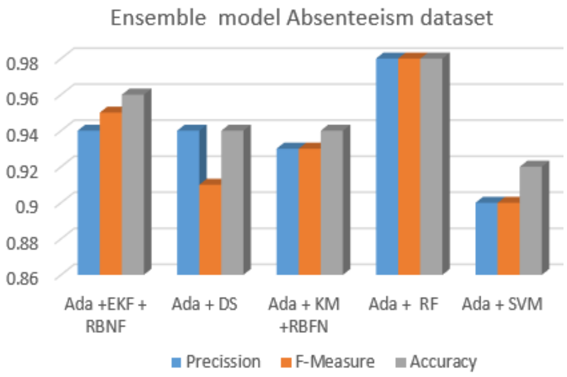


Figure 20 Precision, F-Measure and Accuracy

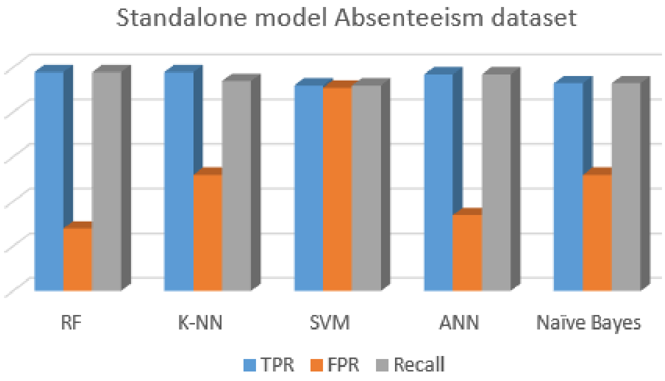


Figure 21 TPR, FPR and Recall

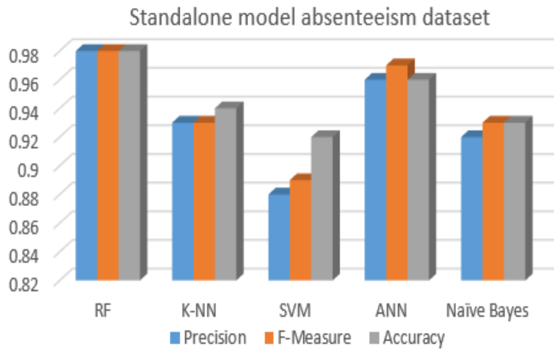


Figure 22 Precision, F-Measure and Accuracy

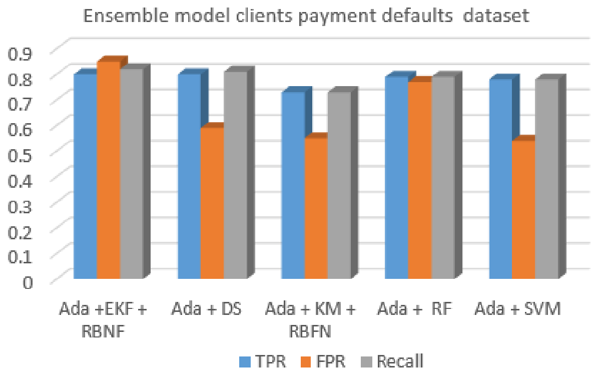


Figure 23 TPR, FPR and Recall

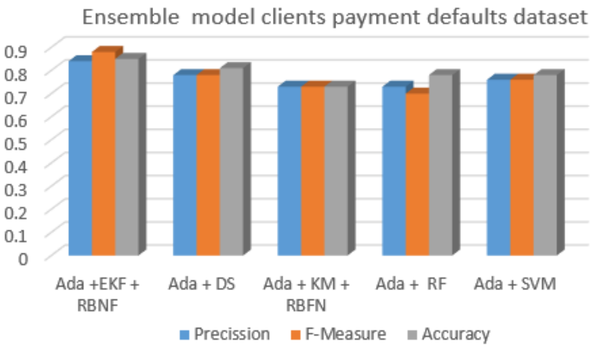


Figure 24 Precision, F-Measure and Accuracy

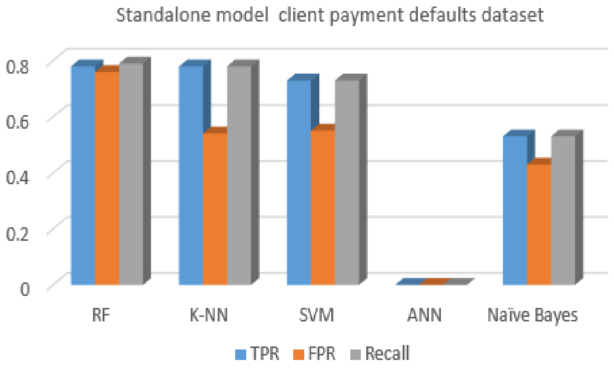


Figure 25TPR, FPR and Recall

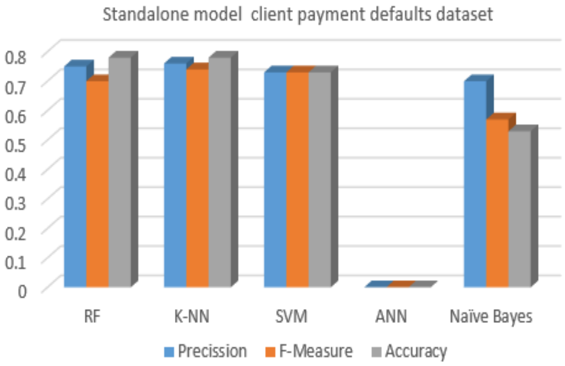


Figure 26 Precision, F-Measure and Accuracy

V Conclusion and Further Work

Even though ensemble algorithms have been widely used extensively in science and engineering applications, nevertheless there is a need for improved prediction accuracy of the algorithm. EKF has been considered as a benchmark algorithm in estimating the state of a system due to its recursive structure, faster convergence and ability to correct itself without storing current or past estimates. Therefore, in this paper we proposed a model that integrates EKF as an optimizing agent to enhance the training parameters of RBFN. Then applied AdaBoost as a meta-algorithm to generate and combine several weak classifiers that produces a stronger predictive output. A performance comparison of the model was carried out using breast cancer survivability, diabetes diagnostic, staff absenteeism and clients credit card payment default datasets that were obtained from the UCI repository. The result shows a good prediction outcome, minimizes overfitting, and improves convergence rates of the model compared with other standard standalone and similar ensemble RBFN models trained with K-means algorithm or Support Vector Machine. Likewise, the prediction accuracy of the ensemble logistic model proposed prosed on cancer dataset is 94% when 70% and 30% of the dataset were used for training and testing the model respectively. We found that the performance of Random Forest as a standalone algorithm or as an ensemble classifier were highly competitive compared with other models used in this study. The findings indicated that using EKF to train RBFN can improve the performance efficiency of ensemble algorithms significantly. The study has gone some way towards improving our knowledge and enhancing prediction accuracy through the unification of EKF, RBFN and AdaBoost algorithms as an ensemble model. The prediction performance of the proposed ensemble logistic regression model also outperforms some of the existing predictive models. In the future, further research will be focused on the application of the proposed models on complex, imbalance datasets, the effects of diversity and algorithmic settings on prediction accuracy, combination methods and possible extension of the ensemble logistic model.

References

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

Adegoke, V., Chen, D. & Banissi, E., 2017. *Prediction of breast cancer survivability using ensemble algorithms.* Osijek, Croatia, IEEE.

Adegoke, V., Chen, D., Banissi, E. & Barikzai, S., 2018. *Enhancing Ensemble Prediction Accuracy of Breast Cancer Survivability and Diabetes Diagnostic using optimized EKF-RBFN trained prototypes.* Porto, Portugal, Springer Verlag.

Adegoke, V. F., Chen, D., Barikzai, S. & Banissi, E., 2017. *Predictive Ensemble Modelling: Experimental Comparison of Boosting Implementation Methods.* UK, IEEE.

Alghamdi, M. et al., 2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PLoS One,* 12(7).

Apreutesei, N. A. et al., 2018. Predictions of ocular changes caused by diabetes in glaucoma patients. *Computer Methods and Programs in Biomedicine,* Volume 154, pp. 183-190.

Barakat, N., Bradley, A. P. & Barakat, M. N. H., 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.

BHF, 2018. *CVD Statistics – BHF UK Factsheet,* s.l.: BHF(British Heart Foundation).

Bychkov, D. et al., 2018. *Deep learning based tissue analysis predicts outcome in colorectal cancer,* s.l.: Scientific Reports.

Cancer Research, U., 2018. *Cancer Research UK.* [Online] Available at: http://www.cancerresearchuk.org/ [Accessed 08 20 2018].

Changxin, L., Jinliang, D. & Tianyou, C., 2014. *Robust Prediction for Quality of Industrial Processes.* Hailar, IEEE.

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

Csank, J. T. & Connolly, J. W., 2016. *Model-Based Engine Control Architecture With an Extended Kalman Filter,* San Diego, California: NASA STI Program.

Dagliati, A. et al., 2018. Machine Learning Methods to Predict Diabetes Complications. *J Diabetes Sci Technol,* 12(2), pp. 295-302.

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).

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

Freud, Y. & Schapire, R. E., 2014. *Boosting Foundations and Algorithms.* s.l.:MIT Press.

Freund, Y. & Schapire, R., 1997. A decision-theoretic generalization of on-line learning and application to boosting. *Journal of computer and system sciences,* Volume 55, pp. 119-139.

Gan, M., Peng, H. & Dong, X.-p., 2012. A hybrid algorithm to optimize RBF network architecture and parameters for nonlinear time series prediction. *Applied Mathematical Modelling,* p. 2911–2919.

Ghosh, J. & Acharya, A., 2011. Cluster ensembles. *WIREs Data Mining Knowl Discov,* p. 305–315.

Haykin, S., 2008. *Neural Networks and Learning Machines: A Comprehensive Foundation.* Third ed. s.l.:Prentice Hall.

Huang, D., Wang, C.-D., Lai, J.-H. & Kwoh, C.-K., 2017. Toward Multi-Diversified Ensemble Clustering of High-Dimensional Data. *Machine Learning - Computer Vision and Pattern Recognition.*

ITV, 2018. *Surge in heart attacks and strokes predicted as diabetes epidemic takes its toll,* London, UK: ITV Report.

Jin, R. & Zhang, J., 2007. Multi-Class Learning by Smoothed Boosting. *Machine Learning,* Volume 67, p. 207–227.

Kalman, R., 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME – Journal of Basic Engineering,* Volume 82, pp. 35-45.

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), pp. 21-32.

Kaur, M., Jat, A. & Ajith, R., 2019. Pediatric Irritable Bowel Syndrome Prediction Using 2 - Tier Ensemble Classifier. *International Journal of Computer Information Systems and Industrial Management Applications.,* 11(2019), pp. 39-45.

Kelwade, J. & Salankar, S., 2016. *Prediction of heart abnormalities using Particle Swarm Optimization in Radial Basis Function Neural network.* s.l., IEEE, pp. 793 - 797.

Kobetski, K. & Sullivan, J., 2015. Improved Boosting Performance by Explicit Handling of Ambiguous Positive Examples. *Pattern Recognition Applications and Methods,* Volume 318.

Kuncheva, L., Hadjitodorov, S. & Todorova, L., 2006. *Experimental Comparison of Cluster Ensemble Methods.* Florence, Italy, IEEE.

Kwon, S. & Lee, S., 2016. Recent Advances in Microwave Imaging for Breast Cancer Detection. *International Journal of Biomedical Imaging.*

Lee, Y., Han, D. & Ko, H., 2013. Reinforced AdaBoost Learning for Object Detection with Local Pattern Representations. *The Scientific World Journal,* Volume 2013, p. 14 pages.

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

Li, T.-C.et al., 2018. Development and validation of prediction models for the risks of diabetes-related hospitalization and in-hospital mortality in patients with type 2 diabetes. *Metabolism,* August, Volume 85, pp. 38-47.

Malathi, P. & Suresh, G., 2014. *Recognition of isolated words of esophageal speech using GMM and gradient descent RBF networks.* s.l., IEEE, pp. 174 - 177.

Mangiafico, S. S., 2015. *An R companion for the handbook of biological statistics.* 1.3.3 ed. New Brunswick, NJ: Rutgers Cooperative Extension.

McDonald, J. H., 2014. *Handbook of Biological Statistics.* 3rd ed. Sparky House Publishing: Sparky House Publishing.

McGinley, B. et al., 2010. Spiking neural networks for breast cancer classification using radar target signatures. *Progress In Electromagnetics Research C,* Volume 17, pp. 74-94.

Mukherjee, I., Rudin, C. & Schapire, R., 2013. The Rate of Convergence of AdaBoost. *Journal of Machine Learning Research,* Volume 14, pp. 2315-2347.

Pak, F., Kanan, H. & Alikhassi, A., 2015. Breast cancer detection and classification in digitalmammography based on Non-SubsampledContourlet Transform (NSCT) and Super Resolution. ComputerMethodsandProgramsin Biomedicin*e,* Volume 122, pp. 89-107.

Puskorious, G. & Feldkamp, L., 1994. Neurocontrol of nonlinear dynamical systems with kalman filter trained recurrent networks. *IEEE Trans. Neural Networks ,* Issue 5, pp. 279-297.

Qiao, X., Chang, W., Zhou, S. & Lu, X., 2016. *A prediction model of hard landing based on RBF neural network with K-means clustering algorithm.* s.l., IEEE, pp. 462-465.

Ribeiro, M., 2004. *Kalman and Extended Kalman Filters: Concept, Derivation and Properties,* s.l.: CiteSeer.

Sapate, S. & Talbar, S., 2016. Mammograms, An Overview of Pectoral Muscle Extraction Algorithms Applied to Digital. *Studies in Computational Intelligence.*

Saravanakumar, S. & Thangaraj, P., 2019. A Computer Aided Diagnosis System for Identifying Alzheimer’s from MRI Scan using Improved Adaboost. *Journal of Medical Systems,* February.

Schapire, R., 2013. Explaining AdaBoost. In: *Empirical Inference.* s.l.:Springer, pp. 37-52.

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

Schapire, R., Freund, Y., Bartlett, P. & Lee, W., 1998. Boosting the Margin:A New Explanation for the Effectiveness of Voting Methods. *The Annals of Statistics,* 26(5), pp. 1651-1686.

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

Society American Cancer, 2015. *American Cancer Society. Breast Cancer Facts & Figures 2015-2016.* [Online]
Available at: https://www.cancer.org/
[Accessed 05 May 2017].

Soni, R., Nigam, M. & Mitra, R., 2015. *Implementation of hybrid neuro-fuzzy controller based on radial basis function network with gradient descent algorithm for non-linear system.* s.l., IEEE, pp. 178 - 183.

Swapna, G., Vinayakumar, R. & K.P, S., 2018. Diabetes detection using deep learning algorithms. *ICT Express,* 2(2), pp. 243-246.

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

Wan, E. & Merwe, R., 2000. The Unscented Kalman Filter for Nonlinear Estimation, Communications, and Control Symposium. *Adaptive Systems for Signal Processing,* pp. 153-158.

Wang, H., Zheng, B., Yoon, S. W. & Ko, H. S., 2018. A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research,* 267(2), pp. 687-699.

Wang, Y. et al., 2015. *Fuzzy radial basis function neural network PID control system for a quadrotor UAV based on particle swarm optimization.* s.l., IEEE, pp. 2580 - 2585.

Weedon-Fekjær, H., Romundstad, P. & Vatten, L., 2014. Modern mammography screening and breast cancermortality: population study. *MMJ,* Volume 348:g3701.

WHO, 2016. *Global report on diabeteS- World Health Organization,* Geneva 27, Switzerland: WHO Library Cataloguing-in-Publication Data.

Xie, W., Li, Y. & Ma, Y., 2015. Breast mass classification in digital mammography based on extreme learning machine. *Neurocomputing,* 73(3), pp. 930-941.

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), pp. 361-371.

Zheng, T. et al., 2017. A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics,* Volume 97, pp. Pages 120-127.

Zou, Q. et al., 2018. Predicting Diabetes Mellitus With Machine Learning Techniques. *National Center for Biotechnology Information - Frontiers in Genetics,* Volume 9:515.