Intelligent Health Risk and Disease Prediction Using Optimized Naive Bayes Classifier

Latifah Alamer[1*], Iman Mohammad Alqahtani[2] and Ebtesam Shadadi[3]

[1*]Information Technology & Security, College of Computer Science & Information Technology, Jazan University, Jazan, Kingdom of Saudi Arabia. laalamer@jazanu.edu.sa

[2]Computer and Information Systems Department, King Khalid University, Abha, Kingdom of Saudi Arabia. ealqahtani@kku.edu.sa

[3]Department of Computer Science, College of Computer Science and Information Technology, Jazan University, Jazan, Kingdom of Saudi Arabia. ashedadi@jazanu.edu.sa

Received: September 09, 2022; Accepted: November 17, 2022; Published: February 28, 2023

Abstract

Machine learning is the subset of Artificial Intelligence and it is used for prediction various real time data analytics applications. Health care monitoring is the major area to analyse the result and make effective decisions. We need intelligent and automated process for predicting diseases using medical dataset. Machine learning methods are proposed to handle the dataset. Smart healthcare prediction is proposed to identify the user or patient information or symptoms as an input. Our system has forecasting accuracy index based on likelihood of the disease and health information. We use Naive bayes classifier algorithm for handling classification, prediction and accuracy index of dataset. Our algorithm measures the disease percentage and train the dataset. Once the prediction result will appears based on effective decision to be taken. In our work, we are taken 20000 train dataset and 7500 test data set for evaluation. TensorFlow simulator is used to simulate the system and measure accuracy. In this system achieves 95% accuracy and performance result compared with existing methods.

Keywords: Machine Learning, Healthcare Dataset, Prediction, Accuracy, Simulation.

1 Introduction

Machine is generative approach for measure prediction based on different instances. It is the AI based approach and promote various machine learn result to process the dataset. In this case we need to process the data, recognising dataset, effective decision making and accuracy prediction [1]. It is the programming approach to process the dataset and optimize the result. Machine learning has two factors such as research model specification and preparation of effective modelling values. The dataset consist of user/patient information, personal details, symptoms and test results. The dataset contains history of information about patients with repository as storage medium [2][3].

Journal of Internet Services and Information Security (JISIS), volume: 13, number: 1 (February), pp. 01-10
DOI: *10.58346/JISIS.2023.I1.001*

*Corresponding author: Information Technology & Security, College of Computer Science & Information Technology, Jazan University, Jazan, Kingdom of Saudi Arabia.

It is strong recommendation system to handle the medical information and address the issue. We need well experienced data processing system to identify the diseases and record it. Machine learning is used for classify the dataset and predict the sign and disease values. In this case each dataset analysed and produce the result. We need automated system to monitor the symptoms and improve the patient health conditions. Various classifiers is available for handling the dataset and classify based volume of information [4].

The patient information are stored in centralized cloud server and each information are stored each patient repository. For each subset values classified based on clinical, administrative [5], academic [6], educational [7] and disease background [8]. Each symptom is also classified from nature disease and past history. In this paper, section 2 describes various related works, section 3 gives proposed methodology and classification, section 4 explains simulations and experimental results and section 5 gives conclusion.

2 Related Works

Mung et al, Plethora Risk health care system is available to handle the dataset and most of the researchers are focused as disease risk. Prediction model is developed based single candidate key diseases from the database. But the major issue is handling non linear and NO SQL dataset [9]. Nowadays huge volume of unstructured data can use for processing. Binary classification is proposed by Wong et al. In this case each disease problems are classified from past medical records and noted as labelled classifier. Due to this case, if patient or user removed or vacate means complete database can be removed and not possible maintain keep record policies [10].

Problem is sorting is single label values, classification issues and object classification. Dinat et al, proposed health records are classified as HoN-Code Dataset principle. Hu et al, proposed EHRS dataset for predicting cancer disease [11]. The surgical information also stored in database patient history monitoring [12]. The mortality rate is calculated as nutrition assessment index with respect to hypertension samples and measures the accuracy. Below table 1 shows that the result of various average morality index [13][14].

Table 1: Hypertension Result of Various Sample Selected from EHRS Result

Dataset	85000 patients and 10000 controls
Factors	10 index features
Birth year	1990 – 2015
Gender	50% male and 50% female
BMI Index	60 – 80 range
Diastolic	30 to 50
Systolic BP	0.45 to 0.75
Triglycerides	1.23 to 2.75
cholesterol	75 to 120
Urine Albumin	100 to 200
Creatinine Ratio	0 to 1

Attari et al, specifies RNA virus is contiguous disease and it is classified by using neural network classifier [15]. Influenza is another disease to handle the human levels such as h1, h3 and h5 features. As per the report said during pandemic the dataset can be classified as human, avian and swine features [16]. It is very difficult to differentiate and categorise the patient. Social media is another platforms to select the features using Twitter and Facebook information.

Today scenario development of social platforms and information exchange platforms provides multi user environment and significant role to analyse the sentiment index [17]. Manikandan et al, proposed a prediction system for handling swine flu using micro blog dataset optimization. Tuberculosis is another diseases and it is predicted based on deep learning. As various researchers inputs we need a classifier for handling healthcare dataset and select the values based on multi object representations. Each classifier can be tested by using Deep convolution neural network and support vector machine is represented for clinical analysis [18].

3 Convolution Neural Classifier – Proposed Model

Machine learning is the method to learn and select the accuracy based on prediction.

Cornerstones: Selection of electronic health record (EHR) information is major key player for our proposed system. In this case we proposed quantitative model for predicting risks.

Research Contributions: Our main objective is to predict the health risk and provide the decision based on medical prescription. User or patient health record can be tested and prevent the disability.

Data driven: In this case the EHR values are classified to into sub groups and clinical decision tree generated based on TensorFlow.

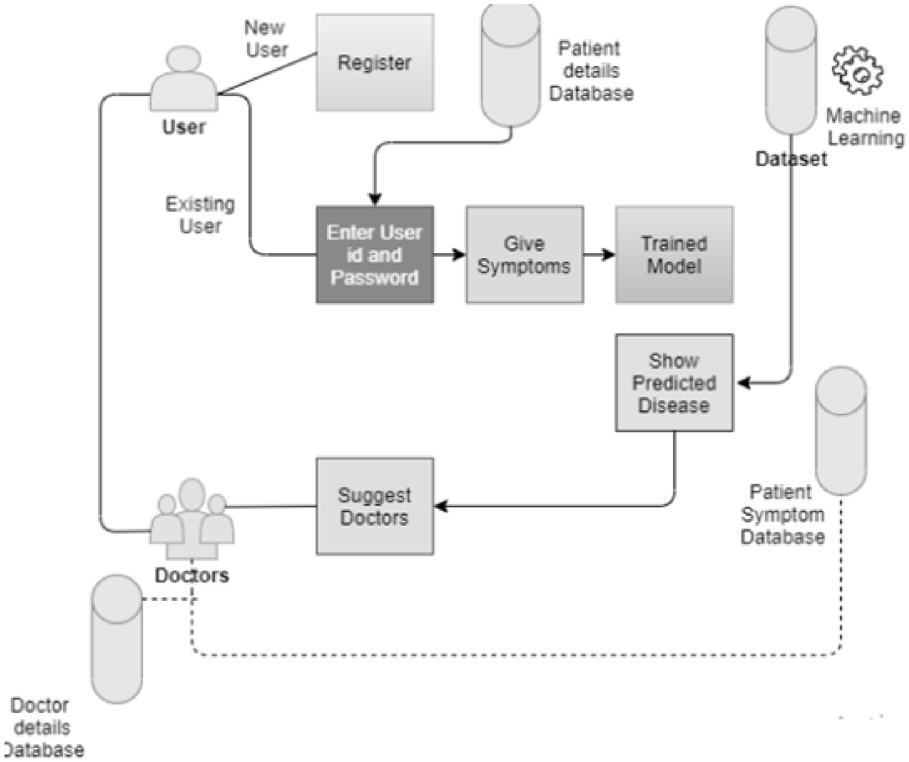


Figure 1: Proposed Electronic Health Record Storage and Classification System

Machine learning is proposed to extract the feature from input dataset and select the values such as forecast, diagnosis, disease, virus feature, decision support values. The main challenges are irrelevant dataset features and biomedical dataset values such as irrelevant data, noisy contents, dirty data and incompleteness. We propose optimized classifier with below contributions.

- Classify the dataset using machine learning - supervised learning features.
- Identify the risk rate based on risk prediction or disease prediction using health record values.
- Identify the feature direction and classify the result.

The representation from Fig.1. Shows that web-based storage forum for predicting disease. The manifest information is recorded based symptoms and conditions. The user or patient will select the feature and analyse the dataset. Based on EHR record our optimized classifier classified the dataset by using below Table 2 representations.

Table 2: Learning Capabilities Based on Dataset

Supervised Learning	Categorise the labelled dataset and handle the structured dataset – Labelled Values: patient information, clinical records
Unsupervised Learning	Unlabelled dataset and map the attributes – Record multimedia contents, NO SQL unstructured dataset
Semi-Supervised Learning	Combination of Labelled and unlabelled dataset – Intelligent Automation and Inference
Reinforcement Learning	Rewards or rating the feature based on machine learning capabilities
Deep Learning	Complexes neural network feature with respect to classified results

4 Optimized Classification – Naive Bayes Classifier

Our proposed classifier classified the labelled / unlabelled dataset to build the deep belief network model using classification, clustering, optimization and regression features.

Table 3: Dataset Classification and Feature Selection

Dataset	20000 train dataset and 7500 test data set
Classification	Grouping the dataset – Quantify the values
Clustering	Derive the Intelligence - Qualitative features and Parameters
Regression	Derive the features and match the data points
Optimization	Attribute function and performance index

Optimized Multiclass Classifier

This method has two level of classifier process

a. Adaption of Algorithm: Select the customized machine learning algorithm to find out labelled learning results

b. Transformation of Problem: Convert multi labelled dataset into single value classifier transfer the result

The below equation describe the Naive Gradient distance calculation,

$$\theta_{t+1} = \theta_t - \mu . \nabla L_t(\theta_t) \tag{1}$$

From which θ is the labeled parameters with weighted values representation as $\nabla L_t(\theta_t)$.

Binary Classification is applied to measure N classifier accuracy by using multi labelled results

$$\left.\begin{aligned} \theta_{t+1} = \ & \theta_t - \mu[(1-v).\nabla L_t(\theta_t) + v.s_{t+1}] \\ s_{t+1} \leftarrow \ & \beta.s_t + (1-\beta).\nabla L_t(\theta_t) \end{aligned}\right\} \tag{2}$$

Selected labelled power set is applied to measure the values of s and t values. It is the complex number factor to set round robin classification. The key pair can be generated as x,y,z. The learning rate is applied as follows.

$$\mu_{t+1} \leftarrow \mu_t + \beta.\nabla L_t(\theta_t).\nabla L_t(\theta_{t-1}) \tag{3}$$

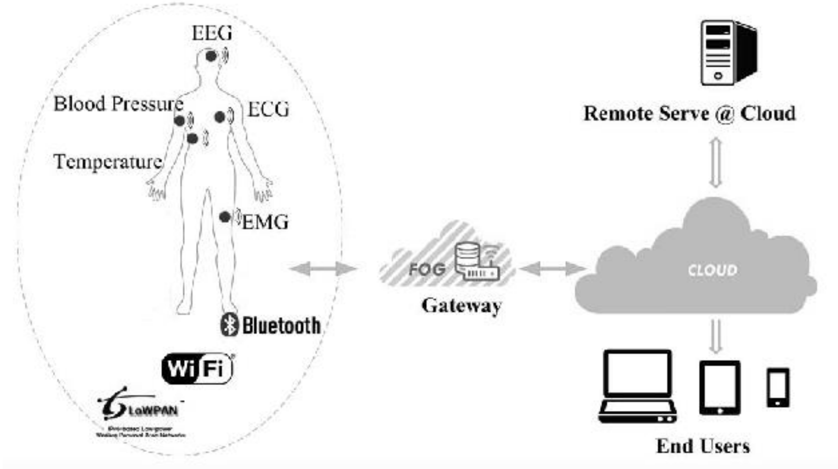


Figure 2: Data Classification Based on User Input Recording System

Optimized Random K-Label set is generated based on labels and size can be verified by k. The threshold value is calculated as follows.

$$\theta_{t+1} \leftarrow \theta_t - \mu_{t+1}\left[\frac{(1-v_1).\nabla L_t(\theta_t)+v_1.f'_{t+1}}{\sqrt{(1-v_2)\left(\nabla L_t(\theta_t)\right)^2+v_2.s'_{t+1}+\epsilon}}\right] \tag{6}$$

$$f_{t+1} \leftarrow \varphi_1.f_t + (1-\varphi_1).\left(\nabla L_t(\theta_t)\right) \tag{7}$$

$$f'_{t+1} \leftarrow (1-\varphi_1^{t+1})^{-1}.g_{t+1} \tag{8}$$

$$\left.\begin{aligned} s_{t+1} &\leftarrow \varphi_2.s_t + (1-\varphi_2)\left(\nabla L_t(\theta_t)\right)^2 \\ s'_{t+1} &\leftarrow (1-\varphi_2^{t+1})^{-1}.s_{t+1} \end{aligned}\right] \tag{9}$$

```
**Algorithm 1:** HER Selection
**Parameters:** set $u,\beta$
$\theta_0 \leftarrow 0$ or random
**for** t ←0 to (no_of iterations-1)  **do**
Compute learning rate
$\mu_{t+1} \leftarrow \mu_t + \beta.\nabla L_t(\theta_t).\nabla L_t(\theta_{t-1})$
 Compute $\theta_{t+1}$
        $\theta_{t+1} \leftarrow \theta_t - \mu_{t+1}[(1-u).\nabla L_t(\theta_t) + u.f_{t+1}]$
$f_{t+1} \leftarrow \beta.f_t + (1-\beta).\nabla L_t(\theta_t)$
**end for**
```

Deep learning is applied to measure and generate the decision tree using Deep Belief Network using Predictive analysis.

The Naive Bayes Classifier is a dynamic method to create the lables and map the dataset. The below Fig 2 shows that the mapping of the object and deep belief network generated model for classifying dataset. The class label can be selected as choices and each feature recorded as range of each representations. This is an average value measured by transmitted information. The number of logs, events, probability values and expectation are recorded. The amount of entry is calculated as

$$HWt = \frac{\sum_{i=1}^{N} D_i(x)}{N} \tag{10}$$

Apply decision tree to separate the data with respect to select positive and negative false positions

$$h_f(x) = \begin{cases} 1 & \sum_{t=1}^{T} (\log 1/\beta_t) h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \log 1/\beta_t \\ 0 & \sum_{t=1}^{T} (\log 1/\beta_t) h_t(x) < \frac{1}{2} \sum_{t=1}^{T} \log 1/\beta_t \end{cases} \tag{11}$$

Algorithm 2: Classifier Rate

```
Parameters: set $\beta$ , $\beta_1$, $\beta_2$,$v_1$,$v_2$ , and $\in$
$\theta_0 \leftarrow 0$ or random
for t←0 to (no_of iterations-1) do
 Compute learning rate
 $\mu_{t+1} \leftarrow \mu_t + \beta.\nabla L_t(\theta_t).\nabla L_t(\theta_{t-1})$
 Compute $\theta_{t+1}$
 $\theta_{t+1} \leftarrow \theta_t - \mu_{t+1}\left[\frac{(1-v_1).\nabla L_t(\theta_t) + v_1.f'_{t+1}}{\sqrt{(1-v_2)(\nabla L_t(\theta_t))^2 + v_2.s'_{t+1}} + \epsilon}\right]$
 $f_{t+1} \leftarrow \varphi_1.f_t + (1-\varphi_1).(\nabla L_t(\theta_t))$
 $f'_{t+1} \leftarrow (1-\varphi_1^{t+1})^{-1}.g_{t+1}$
 $s_{t+1} \leftarrow \varphi_2.s_t + (1-\varphi_2)(\nabla L_t(\theta_t))^2$
 $s'_{t+1} \leftarrow (1-\varphi_2^{t+1})^{-1}.s_{t+1}$

end for
```

It is a general method to identify the patient dataset and apply medical data feature. Here the exponential rate, processing existing feature and accuracy factors are noticed. In this case disease prediction can be recorded based on Naives results. The accuracy is calculated by using membership value and entropy count.

5 Experimental Setup

Input: (X, Y) ≡ X:Y 512 X 512 X 3 Layers

Output: Optimized Naive Bayes Classifier - Accuracy

Step 1: Deep Q based Data Classification - Index

Step 2: Calculate Bayes (X, Y) = (1/N) ∑(i=0) ^(X-1) G 〖X(i)- ∑(i-0) ^(Y-1) X [Y(i)- Y (min X)]]

Step 3: Accuracy (X, Y) = ∑(i=0) ^(X-1) G / ((V(i)-S(i)/N))

Step 4: Record the store value and repeat the process until n<0

The healthcare dataset is taken from Smart Tenz dataset is selected from healthcare industries and Google TensorFlow is used to simulate the system using below conditions.

Table 4: TensorFlow Input Dataset feature

Train and Test Dataset (X and Y)	Optimized Naive Bayes Classifier
Input values	512 X 512 X 3 Layers
Connected Points	3 connected layers
Hidden Node	8,16,32,64
GPU	3.75Ghz Deep Decision tree model
Dimensionality	100 X 100
Dataset	20000 train dataset and 7500 test datasets

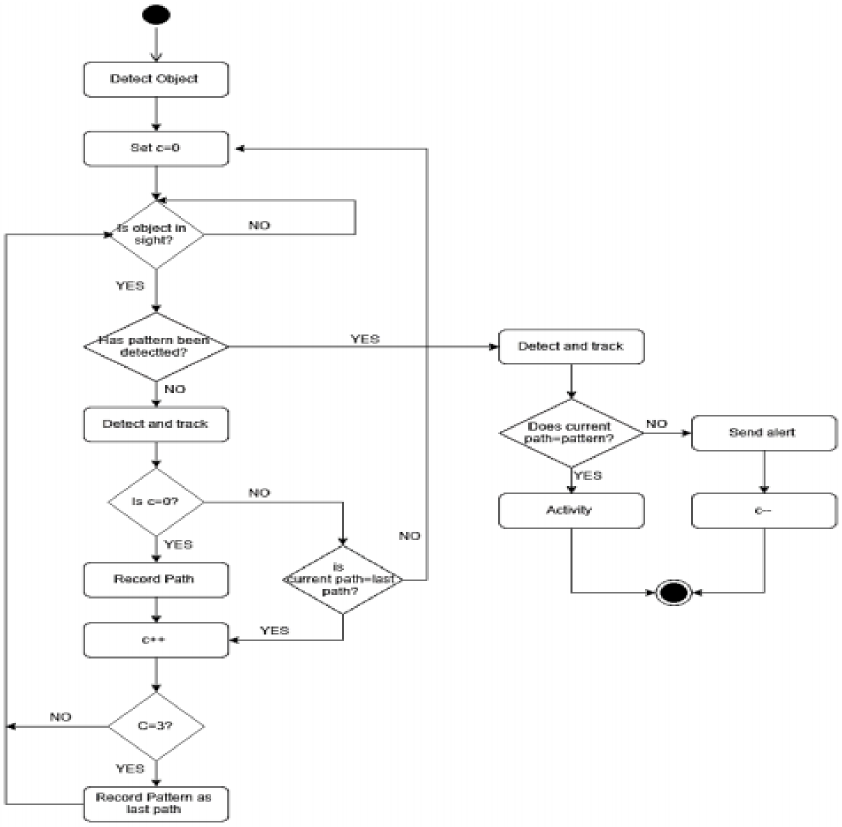


Figure 3: TensorFlow Simulation Flow Generated from TensorFlow

The above Fig 4 is Deep Belief Network for proposed method and collects the layered records. Fig 5 shows that 100 X 100 dimensionality reduction result of optimal resource provisioned results.

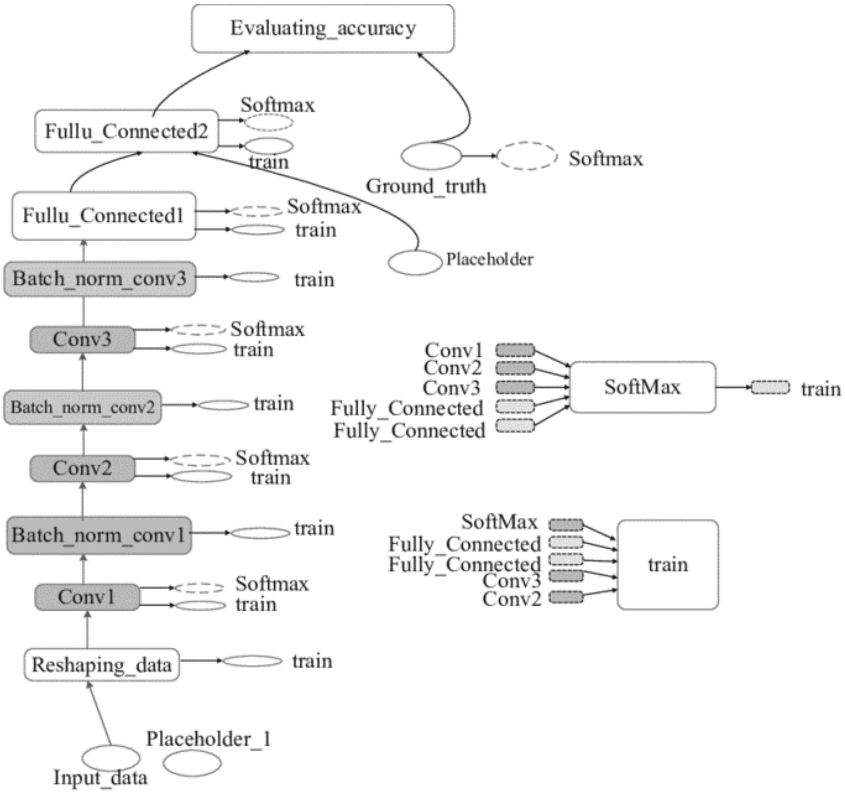
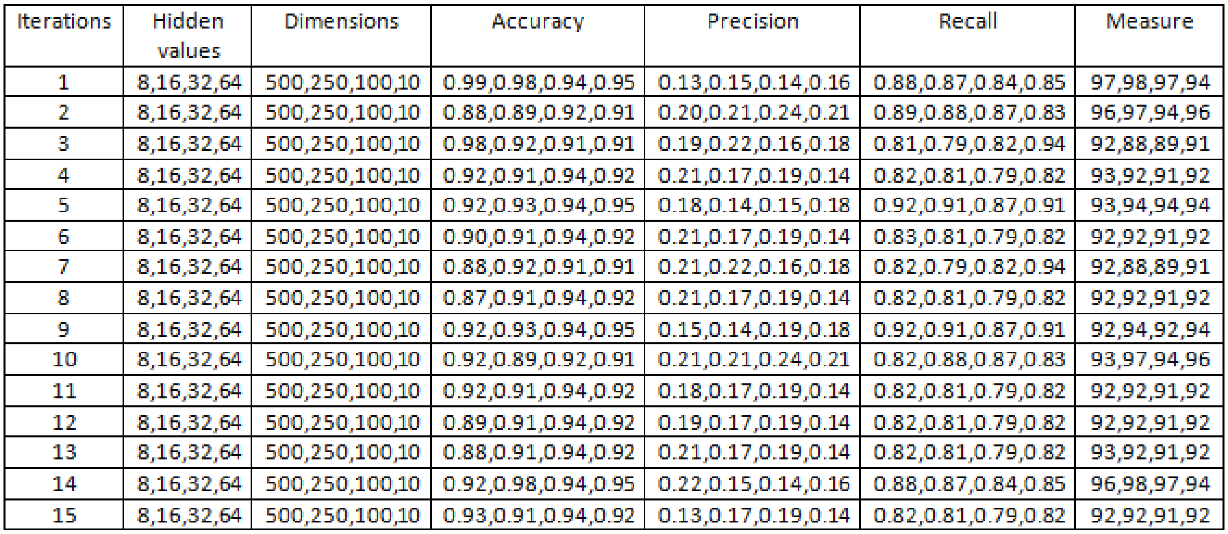


Figure 4: Deep Belief Network Generated based on Input – Naive Bayes result

Table 5: Dataset table derived from TensorFlow and Accuracy Index Prediction (GPU: 3.75Ghz Deep Decision tree model)



Iterations	Hidden values	Dimensions	Accuracy	Precision	Recall	Measure
1	8,16,32,64	500,250,100,10	0.99,0.98,0.94,0.95	0.13,0.15,0.14,0.16	0.88,0.87,0.84,0.85	97,98,97,94
2	8,16,32,64	500,250,100,10	0.88,0.89,0.92,0.91	0.20,0.21,0.24,0.21	0.89,0.88,0.87,0.83	96,97,94,96
3	8,16,32,64	500,250,100,10	0.98,0.92,0.91,0.91	0.19,0.22,0.16,0.18	0.81,0.79,0.82,0.94	92,88,89,91
4	8,16,32,64	500,250,100,10	0.92,0.91,0.94,0.92	0.21,0.17,0.19,0.14	0.82,0.81,0.79,0.82	93,92,91,92
5	8,16,32,64	500,250,100,10	0.92,0.93,0.94,0.95	0.18,0.14,0.15,0.18	0.92,0.91,0.87,0.91	93,94,94,94
6	8,16,32,64	500,250,100,10	0.90,0.91,0.94,0.92	0.21,0.17,0.19,0.14	0.83,0.81,0.79,0.82	92,92,91,92
7	8,16,32,64	500,250,100,10	0.88,0.92,0.91,0.91	0.21,0.22,0.16,0.18	0.82,0.79,0.82,0.94	92,88,89,91
8	8,16,32,64	500,250,100,10	0.87,0.91,0.94,0.92	0.21,0.17,0.19,0.14	0.82,0.81,0.79,0.82	92,92,91,92
9	8,16,32,64	500,250,100,10	0.92,0.93,0.94,0.95	0.15,0.14,0.19,0.18	0.92,0.91,0.87,0.91	92,94,92,94
10	8,16,32,64	500,250,100,10	0.92,0.89,0.92,0.91	0.21,0.21,0.24,0.21	0.82,0.88,0.87,0.83	93,97,94,96
11	8,16,32,64	500,250,100,10	0.92,0.91,0.94,0.92	0.18,0.17,0.19,0.14	0.82,0.81,0.79,0.82	92,92,91,92
12	8,16,32,64	500,250,100,10	0.89,0.91,0.94,0.92	0.19,0.17,0.19,0.14	0.82,0.81,0.79,0.82	92,92,91,92
13	8,16,32,64	500,250,100,10	0.88,0.91,0.94,0.92	0.21,0.17,0.19,0.14	0.82,0.81,0.79,0.82	93,92,91,92
14	8,16,32,64	500,250,100,10	0.92,0.98,0.94,0.95	0.22,0.15,0.14,0.16	0.88,0.87,0.84,0.85	96,98,97,94
15	8,16,32,64	500,250,100,10	0.93,0.91,0.94,0.92	0.13,0.17,0.19,0.14	0.82,0.81,0.79,0.82	92,92,91,92

Table.5 shows the accuracy index of 20000 trained data and 7500 test dataset and the average accuracy index measure as 95%. From this result our proposed system is compared with existing medical dataset prediction method.

Table 6: Comparison of Proposed Method with Existing Dataset

Methods	Model	Index Accuracy	Dimensions
Semantic_Net	Dataset Classification	78%	512 X 512 X 3 Layers
Machine_Vision	Learning Agent Modeling	82%	512 X 512 X 3 Layers
COLAB	Agent Model	84%	512 X 512 X 3 Layers
ML_SQL	NO SQL Dataset	85%	512 X 512 X 3 Layers
Optimized Naive Bayes Classifier	Machine Learning	95%	512 X 512 X 3 Layers

From the Table 6 comparison of proposed method with existing model with same dimension features. Compare with that our proposed system gives more accuracy index.

6 Conclusion

Various applications are currently available for handling medical or healthcare dataset. It is electronically storage system to handle the dataset and classification. The major problem is prediction diseases accuracy. We proposed Optimized Naive Base classifier for predicting diseases accuracy using machine learning. TensorFlow simulator is used to simulate system and input dataset is taken for deep belief network. The Deep belief network is generated and measures the accuracy index. Based on that our proposed system has achieved 95% accuracy index and we compared with existing method also. In future we propose various classifiers to predict particular disease classification.

References

[1] Naveenkumar, S., Kirubhakaran, R., Jeeva, G., Shobana, M., & Sangeetha, K. (2021). Smart health prediction using machine learning. *International Research Journal on Advanced Science Hub (IRJASH), 3*(3), 124-128.

[2] Shinde, S.A., & Rajeswari, P.R. (2018). Intelligent health risk prediction systems using machine learning: a review. *International Journal of Engineering & Technology*, *7*(3), 1019-1023.

[3] Ahamed, M.I., & Kumar, K.S. (2019). Modelling of electronic and optical properties of Cu2SnS3 quantum dots for optoelectronics applications. *Materials Science-Poland, 37*, 108-115.

[4] Sun, Y. (2019). The neural network of one-dimensional convolution-an example of the diagnosis of diabetic retinopathy. *IEEE Access*, *7*, 69657-69666.

[5] Manikandan, S., Dhanalakshmi, P., Priya, S., & Teena, A.M.O. (2021). Intelligent and deep learning collaborative method for E-learning educational platform using TensorFlow. *Turkish Journal of Computer and Mathematics Education*, *12*(10), 2669-2676.

[6] Chen, W., Yang, B., Li, J., & Wang, J. (2020). An approach to detecting diabetic retinopathy based on integrated shallow convolutional neural networks. *IEEE Access*, *8*, 178552-178562.

[7] Shorten, C., & Khoshgoftaar, T.M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1-48.

[8] Buda, M., Maki, A., & Mazurowski, M.A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, *106*, 249-259.

[9] Zhang, J., & Mitliagkas, I. (2017). Yellowfin and the art of momentum tuning. *arXiv preprint arXiv:1706.03471*.

[10] Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., & Klein, J.C. (2014). Feedback on a publicly distributed image database: the Messidor database. *Image Analysis & Stereology*, *33*(3), 231-234.

[11] “MESSIDOR dataset,” [Online]. Available: https://www.adcis.net/en/third-party/messidor/.

[12] Garnett, R., Huegerich, T., Chui, C., & He, W. (2005). A universal noise removal algorithm with an impulse detector. *IEEE Transactions on image processing*, *14*(11), 1747-1754.

[13] Ahamed, M.I., & Kumar, K.S. (2019). Studies on Cu2SnS3 quantum dots for O-band wavelength detection. *Materials Science-Poland*, *37*, 225-229.

[14] Irshad Ahamed, Mansoor Ahamed, Sathish Kumar, K., & A. Sivaranjani. (2022). Comparative Energy Bandgap Analysis of Zinc and Tin Based Chalcogenide Quantum Dots. *Revista Mexicana de Física*, *68*(4), 1-8.

[15] Fortino, G., Savaglio, C., Spezzano, G., & Zhou, M. (2020). Internet of things as system of systems: A review of methodologies, frameworks, platforms, and tools. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *51*(1), 223-236.

[16] Ahameda, M.I., Ahamedb, M., Sivaranjanic, A., & Chockalingamd, S. (2021). Energy bandgap studies on copper chalcogenide semiconductor nanostructures using cohesive energy. *Chalcogenide Letters*, *18*(5), 245-253.

[17] Zhang, G., Pan, J., Zhang, Z., Zhang, H., Xing, C., Sun, B., & Li, M. (2021). Hybrid graph convolutional network for semi-supervised retinal image classification. *IEEE Access*, *9*, 35778-35789.

[18] Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., & Kang, H. (2019). Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, *501*, 511-522.

[19] Bakhtina, M., & Matulevicius, R. (2022). Information Security Risks Analysis and Assessment in the Passenger-Autonomous Vehicle Interaction. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 13*(1), 87-111.