A Promising Prediction of Diabetes Using a Deep Learning Approach

Conference Paper · January 2022

DOI: 10.1109/ICCMC53470.2022.9753763

6 authors, including:

Rashiduzzaman Shakil
Daffodil International University
15 PUBLICATIONS **100** CITATIONS
SEE PROFILE

Bonna Akter
Daffodil International University
15 PUBLICATIONS **100** CITATIONS
SEE PROFILE

Fahad Faisal
Daffodil International University
32 PUBLICATIONS **202** CITATIONS
SEE PROFILE

Tahmid Rashik Chowdhury
Islamic University of Technology
5 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Proceedings of the Sixth International Conference on Computing Methodologies and Communication (ICCMC 2022)
DVD Part Number: CFP22K25-DVD; ISBN: 978-1-6654-1027-4

A Promising Prediction of Diabetes Using a Deep Learning Approach

Rashiduzzaman Shakil
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
rashiduzzaman15-2655@diu.edu.bd

Bonna Akter
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
bonna15-2585@diu.edu.bd

Fahad Faisal
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
fahad.cse@diu.edu.bd

Tahmid Rashik Chowdhury
Dept. of CSE
Islamic University of Technology
Dhaka, Bangladesh
eclipse15051@gmail.com

Tonmoy Roy
Dept. of CSE
North South University
Dhaka, Bangladesh
tonmoy5994@gmail.com

Ankit Khater*
Dept. of CSE
Jadavpur University
Jadavpur, India
ankitkhaterak@gmail.com

***Abstract*—Diabetes is a collection of metabolic illnesses caused by a persistently high blood sugar level. If a reliable estimation is achievable, the risk factor and severity of diabetes can be considerably decreased. Because of the low number of labelled data and the prevalence of outliers (or missing values) in diabetes datasets, a robust and reliable diabetes prediction is challenging. Alongside, the incidence rates of diabetes are rising alarmingly every year. Consequently, an early diagnosis of diabetes would be the most crucial step for receiving proper treatment. Hence, a deep learning-based reorganization system has gained popularity regarding disease identification. In this work, we used an updated Convolution Neural Network (CNN) model, modifying different hyper-parameters and layer topologies on the UCI 130 USA Hospitals diabetes dataset. Additionally, five different types of optimizer, namely adaptive moment estimation (ADAM), ADAMAX, Root Mean Square Propagation algorithm (RMSprop), stochastic gradient descent (SGD), and Nesterov accelerated adaptive moment (NADAM), have been analyzed to show a more robust solution. Furthermore, improved accuracy of 99.98% was received by the ADAMAX optimizer.**

***Index Terms*—Diabetes, Deep Learning, Convolution Neural Network, ADAM(Adaptive moment estimation), SGD(Stochastic gradient descent)**

I. Introduction

Diabetes is a very well known disease all over the world. Diabetes is a chronic disease. It occurs when blood glucose crosses the ascertained level and the pancreas isn't able to produce enough insulin for the body. There are two types of diabetes. Those are:

- Type-1 diabetes
- Type-2 or Mellitus diabetes

Among these two types of diabetes, type-2 diabetes [1] creates more complications in the human body. In 2017, approximately 462 million individuals were affected by type 2 diabetes, amounting to 6.28% of the world's population (4.4% of those aged 15–49 years, 15% of those aged 50–69, and 22% of those aged 70+), or a prevalence rate of 6059 cases per 100,000 [2]. Nowadays, it's become a cause for panic worldwide. It's a challenge for the global healthcare community. Hence, diabetes is the ninth reason for early death in 2021 mentioned by WHO [3]. The premature death rate from diabetes is increasing day by day. Between 2000 and 2016, this increased by 5% [4]. The prevalence of diabetes has been rising rapidly in low and middle-income countries. In 2019, an estimated 1.5 million deaths were directly caused by diabetes [3]. Diabetes is also a risk factor for many health complications. It damages the heart, blood vessels, eyes, kidneys, and many other organs of the human body. Its reduced blood flow and increase the risk of a heart attack. Diabetes retinopathy is a major problem for diabetes patients. It's recurrent blindness. Diabetes occurs in 2.6% of global blindness [4]. If this situation continues to occur, it will be a curse globally. Early prediction and diagenesis can be possible through a doctor's knowledge and experience. But there will be a problem of not being able to predict with accuracy and suitability. Early prediction of diabetes can decrease mortality and many other health complications that occur because of diabetes [5]. Many data mining and machine learning methods have been proposed by many researchers for the early prediction of diabetes. But in data mining, there are many hidden layers. So, the accuracy of the proposed method is questionable. Here, we introduced a deep learning-based method for predicting diabetes. The major motive behind this proposed method is to predict diabetes. In our method, we use Diabetes 130-US hospitals for the years 1999-2008 as our data set. It includes the admission of 69980 patients and contains 100,000 records. We fit missing data from the acquired 48 attributes. In this method, we used many deep learning classifiers. such as MLP, DBN, LSTM, and RBFN classifiers for getting accurate predictions. Early prediction of diabetes can help to reduce worldwide premature mortality caused by diabetes. Key contribution of this research paper as follows:

- For detecting diabetic patients, this study intoduced a novel experiment based on CNN architecture with ADAMAX, ADAM, RMSprop, SGD, and NADAM.
- This method automatically recognizes patients who are afflicted.
- This paper introduces the Diabetes 130-US hospitals for years 1999-2008 Data Set from UCI machine learning repository.

II. LITERATURE REVIEW

Very few studies have been performed to predict human disease. Moreover, some enthusiasm has been operated upon to predict disease, prevent disease, predict the risk factor of disease and So on. In this portion, we will retry some previous studies to prove the concept of deep learning and machine learning usability in the driving prediction model, mainly for diabetes. T. Yang et al. [2] presented a deep learning frame work to establish a personalized glucose prediction model for a new subject with type 2 diabetes. They used DTW and a network based deep transfer learning method in the case of improved generalization capability, which obtained 99.57% accuracy with a 5183 pointer for euglycemia and 86.50% for hyperglycemia. S. K. Mohapatra et al. [6] proposed a diabetes detection method based on MLP. In their work, they use the PIMA Indian dataset and use 70% of the data to train the model and another 30% for testing purposes. They obtained 77.5% average accuracy. This dataset contained the data of 768 patients. When this method is used for large datasets, the accuracy can be decreased. M. Alam et al. [7] provide an automated OCTA for recognizing diabetic retinopathy using CNN and VGG16. They utilized two datasets. They utilized a dataset of 52 clinical OCT scans and obtained cross validation performance that revealed 87.65% accuracy and an external validation accuracy of 70.83%. H. Naz and H. Ahuja [8] provided a deep learning technique for the prediction of diabetes. They used the PIMA dataset and applied it to a DL approach. They elicited 98.07% accuracy for the DL and DT classifier algorithms. A deep neural network based prediction approach for predicting diabetes retinopathy from risk factors was referred to by G. Alfian et al [9]. In their referred system, they used a combined DNN+RFE model and obtained 82.033% accuracy. But the dataset they used was small and for a specific population. A system for detecting diabetes using deep learning algorithms was developed by S.KP et al. [10]. They used HRV data from ECG signals. For extracting features, the hybrid network (CNN-LSTM) with SVM obtained 95.7% accuracy. The accuracy result would be an anomaly by extracting dynamic characteristics from the input data. A deep learning based detection method was proposed by M. Rahman et al. [11] for the prediction of diabetes. They used the PIMA Indian dataset. For classification, they used CNN, T-LSTM, and CNN-LSTM for classification. For finding optimal parameters, they used a greedy search algorithm and, in the case of significant features, extraction work on the Boruta algorithm. They obtained 91.38% accuracy on Conv-LSTM and increased this to 97.26% accuracy using cross-validation in Conv-LSTM. P. Roy et al. [12] proposed a deep learning technique for diabetes mellitus prediction. They used the PIMA Indian dataset. And for classification purposes, they used DL, CNN, RBMs, DBNs, AEs, and RNNs classifiers. In this model, they get 91.26% accuracy in the DBN classifier, which is higher accuracy, and the overall average accuracy is 87.27%. T. Vijayakumar et al. [13] worked on Fusion based Feature Extraction Analysis of ECG.(Electro Cardiogram) Signal Interpretation in a Systematic Approach. Their selected dataset contains 350 samples to analyze for train and test. In their existing work, a Pan-Tomkins algorithm was used to detect QRS complexes, and 96.50% accuracy was gained in their work. The application of Virtual reality technology in the online retail industry, as well as future potential, were discussed a variety of situations, from real estate to apparel and cosmetics. Early prediction of coronary artery disease by machine learning was studied by P. Hengjinda and J. I. Z. Chen [14]. Although there is no predefined size for the dataset, 70% and 30% of the data were used for training and testing sequentially. For existing work, SVM (support vector machine) and Nave Bayes algorithm were chosen. SVM outperformed Nave Bayes when compared to both of them. A. Sathesh [15] presented a patient management system based on the Internet of Things (IoT). The Artificial Intelligence idea is employed here to quickly diagnose health issues. The suggested system uses a fuzzy regression approach to handle data between cloud storage as well as the application layer. The AWS algorithm outperforms the FRR, MAS, and ODA algorithms in terms of performance. A. Sungheetha [16] suggested using a data driven model to implement a COVID-19 decision making strategy for risk minimization. They divided the population into ten different segments for evaluating results and finding out the best COVID-19 possibilities based on these disunited parts. A machine learning approach was proposed by A. Sharma et al. [17] to predict diabetic disorders. There were 768 instances and nine characteristics in all. It was worked on by five different machine learning algorithms, with logistic regression achieving the best accuracy, at 80.42 per cent.

III. METHODOLOGY

A. Overview of the Proposed Model

Detection of diabetes is complicated by following several steps. Data is the most important thing for research work. Input data is considered raw material. In this research work, we used 130 US hospital datasets. It represents 10 years of hospital clinical rate data from 1999 to 2008. There are always the same 100000 instances and a total of 55 attributes. We collected data from the UCI machine learning repository [18]. It is a data mining technique that transforms raw data into an understandable format [19]. As this is a huge dataset, most can probably carry missing values easily. Using a simple imputation technique, we handled the missing values. In the preprocessing step, we also use standardization [20] and split the dataset into training and testing datasets. We take 70% of the data for training purposes and another 30% of the data to test the model. CNN means ”convolutional neural network

[21]. It is a multilayered neural network with a particular architecture to detect complicated elements in data. Optimizers or procedures can change the characteristics of neural network. In convolutional neural networks, there are several types of optimizers like ADAM, SGD, RMSprop, Adadelta, Adagrad, and NADAM [22]. We chose to choose optimizers for our model. These are ADAMAX, ADAM, SGD, RMSprop, and NADAM. Among them, we get the best results using the ADAMAX Optimizer. An overall description of each step is presented in Fig. 1.

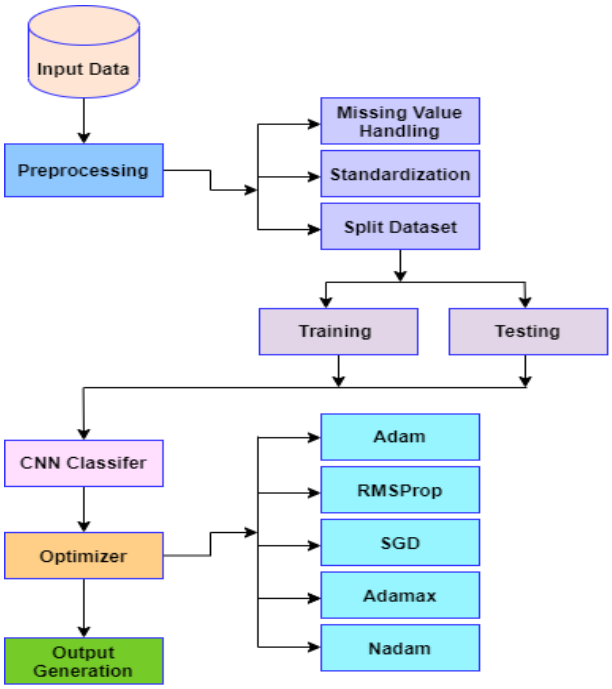

Fig. 1. Working procedure of our model.

B. Optimizer Description

1) ADAM(Adaptive moment estimation): At present time deep learning generally necessitates a significant amount of time and computing resources for training, the optimization technique is the ADAM touchy subject. ADAM optimizer occupies fewer resources and makes the model converge faster [23], which can accelerate the learning speed and improve the effect. ADAM considers as the first order optimization algorithm that can replace the traditional stochastic gradient descent process. In ADAM optimizer learning rate of every single parameter is dynamically adjusted by using the first and second moment estimation of the gradient. In order to makes parameters relatively stable bias connection was also added.

2) RMSprop(Root mean square propagation): RMSProp optimizer considers as the gradient descent with momentum. This optimizer limits the oscillations in the upright direction. That's why any algorithm can increase the learning rate and take horizontal direction converging quickly.

3) SGD(Stochastic Gradient Descent): Stochastic represent the task that's associated with random possibility. By using SGD we select a few samples randomly from the data set.

4) ADAMAX(Adaptive Max Pooling): ADAMAX optimizer is the development form of ADAM. The modification is in the use of infinity norm (ut).

5) NADAM(Nesterov-accelerated Adaptive Moment Estimation): NADAM is constructed with a combination of ADAM and Nesterov Accelerated Gradient(NAG) [23]. NADAM is better than ADAM because NAG could achieve better results than classic momentum.

C. Performance Measure Indices

For the performance measurement of the applied classifier, we have considered several performance metrics named accuracy, sensitivity, false negative rate (FNR), and recall. Mean absolute error [24], Root mean square error [25]. The following equation is used to measure the performance of the classifier. Where, P = Positive class, N = Negative class, TP = True positive, TN = True Negative, FP = False Positive, FN = False Negative.

$$Accuracy = (\frac{TP + TN}{TP + FP + TN + FN}) \times 100\% \quad (1)$$

$$Recall = (\frac{TP}{TP + FN}) \times 100\% \quad (2)$$

$$FNR = (\frac{FN}{TP + FN}) \times 100\% \quad (3)$$

IV. RESULTS AND DISCUSSIONS

A. Our Proposed Deep Learning Classifier

In Fig. 2, we show the accuracy of the models we created using the CNN classifier [26] and ADAMAX optimizer for various dense layers. We used 101766 instances as inputs in our model. For our model, we employed sigmoid activation [27]. We utilize a dense value of 256 in Dense 1 2, which is 128 in Dense 3 and 16 in Dense 4. The function of the Activation Layer is to apply ReLU to the dense layer, and the dropout value for each dense layer was set to 0.25. For each dense layer, we utilized a separate kernel regularizer value of 0.001 for dense 12 and 0.003 for dense 3 & 4.

B. Experimental Outputs Among Different Methods:

In this research work, outcomes were achieved by applying the CNN classifier. And in the CNN classifier, we evaluate the results for five different optimizers. Finally, we find out the average, precision, recall, and F1-score values [28-29] for all of the optimizers. We get the highest accuracy of 99.98%. Here we calculate the average, precision, recall and F1-score for ADAMAX, ADAM, RMSprop, SGD, and NADAM optimizers. In the average value, you can also calculate the macro average and weighted average. Performance metrics are defined as figures and data that are representative of an organization's actions, abilities, and overall quality. We evaluate testing accuracy, sensitivity, false positive, false negative, mean absolute error, mean square error, and root means square error. Here we get the testing accuracy for ADAMAX was 99.98%, ADAM gained 99.71%, RMSprop, SGD, NADAM testing Accuracy gradually increased to 99.59%, 99.91%, and

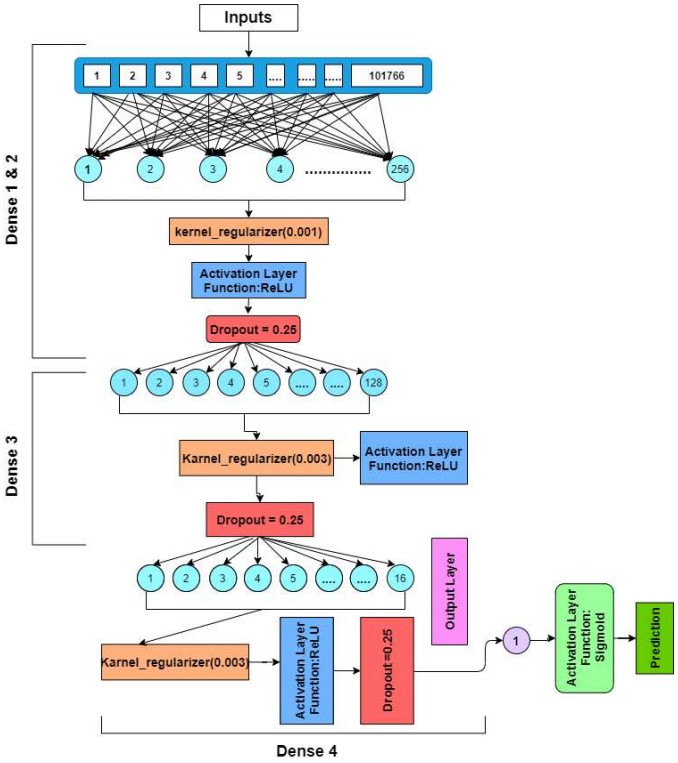

Fig. 2. Visualization functionalities of all optimizers.

99.97%. The sensitivity matrix determines a model's ability to predict true positives in each category. Here is the recall for all optimizers. The values are gradually ADAMAX (99.93%), ADAM (98.80%), RMSprop (99.30%), SGD (99.65%) and lastly, NADAM (99.87%). False negative values occur if your model incorrectly predicts the positive class. The false negative values are ADAMAX (0.06%), ADAM (1.19%), RMSprop (1.69%), SGD (0.34%) and the last one, NADAM (0.12%). From Table 1, we showed all the performance values in one table, step by step.

TABLE I
RESULTS OF THE INTRODUCED ALGORITHM ON THE INTRODUCED OPTIMIZERS

Optimizer	TA	Recall	FNR	MSE	ghoash
ADAMAX	99.98%	99.93%	0.06%	0.0023	0.0134
ADAM	99.71%	98.80%	1.19%	0.0310	0.0750
RMSprop	99.59%	98.30%	1.69%	0.0123	0.0640
SGD	99.91%	99.65%	0.34%	0.0044	0.0294
NADAM	99.97%	99.87%	0.12%	0.0147	0.0335

Figure 03 shown the training accuracy of ADAMAX, ADAM, RMSprop, SGD, and NADAM in a graphical form.The ADAMAX optimizer was the most accurate of all, with a 99.98% accuracy score. RMSprop, on the other hand, was the least accurate, with a 99.59 percent accuracy. The recall is a metric for how well our model detects true positives. As a result, the recall informs us how many individuals with

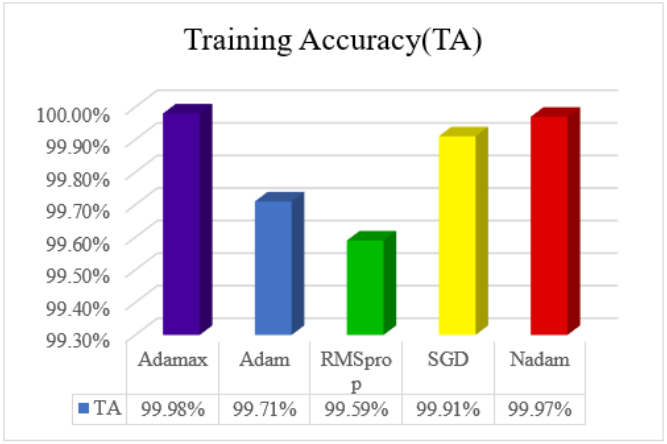

Fig. 3. Graphical representation of Training Accuracy.

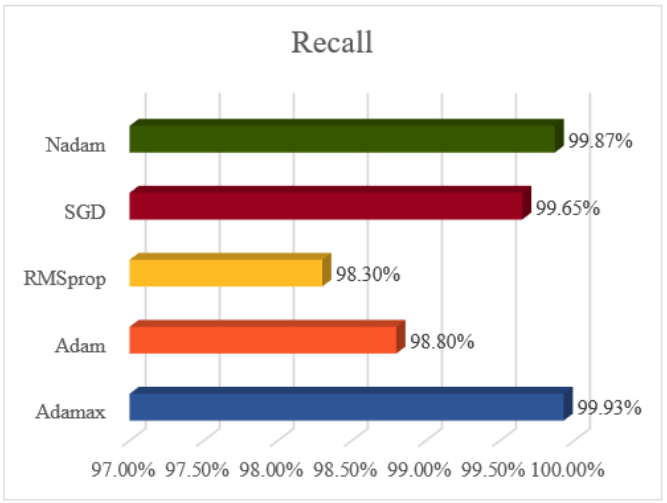

Fig. 4. Recall results for five optimizers.

diabetes we properly recognized. The recall result of our chosen optimizers is shown in Figure 4. The value of the ADAMAX optimizer is 99.53 percent, which is the best among all the optimizers.

V. CONCLUSION

A proposed system for detection of diabetics disease has been Presented in this paper. A total of almost 100000 data are utilized to accomplish this work. For data preprocessing technique handling missing value, Data standardization, and splitting dataset into training and Testing set. For segment the diabetic patient from the dataset CNN classifier is applied. To evaluate the performance of classifier we calculate seven performance metrics. For getting better accuracy we used five optimizer and generate highest accuracy with ADAMAX optimizer , which is 99.98% and the lowest accuracy contained RMsprop optimizer which obtained 99.59% accuracy. In future we will apply more feature extraction technique for segmentation of dataset. And also use more classifier for getting better prediction result in term of diabetics prediction.If we applied

feature selection technique then the overall accuracy may be improved. Our research work introduced specific region people. In future, we will explore on multiple regions and compare with this existing work.

REFERENCES

[1] M. A. B. Khan, M. J. K. Hashim, J. K. King,R. D. Govender, H. Mustafa and J. Al. Kaabi "Epidemiology of type 2 diabetes–global burden of disease and forecasted trends", Journal of epidemiology and global health, vol. 10, no.1, pp.107. 2020.

[2] X. Yu, I. Yang, J. Lu, Y. Shen, W. Lu, W. Zhu, Y. Bao, H. Li, and J. Zhou "Deep transfer learning: a novel glucose prediction framework for new subjects with type 2 diabetes", Complex Intelligent Systems, pp.1-13, 2021.

[3] "The top 10 causes of death," [online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. [Accessed on: 22 – 10 – 2021]

[4] "The Key Facts of Diabetes" [online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed on: 22 – 10 – 2021]

[5] S. Manlangit, S. Azam, and B. Shanmugam "Novel machine learning approach for analyzing anonymous credit card fraud patterns", International Journal of Electronic Commerce Studies, vol. 10, no.2, pp.175-202. 2019.

[6] S. K. Mohapatra, J. K. Swain, and M. N. Mohanty. "Detection of diabetes using multilayer perceptro", International conference on intelligent computing and applications. Springer, Singapore, pp. 109-116, 2019.

[7] D. Le, M. Alam, C. K. Yao, J. I. Lim, Y. T. Hsieh, R. V. Chan, D. Toslak and X. Yao "Transfer learning for automated OCTA detection of diabetic retinopathy", Translational Vision Science Technology, vol. 9, no.2, pp.35-35. 2020.

[8] H. Naz, and S. Ahuja. "Deep learning approach for diabetes prediction using PIMA Indian dataset." Journal of Diabetes Metabolic Disorders, vol. 19, no.1, pp. 391-403, 2020

[9] G. Alfian, M. Syafrudin, N. L. Fitriyani, M. Anshari, P. Stasa, J. Svub and J. Rhee "Deep neural network for predicting diabetic retinopathy from risk factors". Mathematics, vol. 8, no.9, pp.1620, 2020.

[10] G. Swapna, R. Vinayakumar and K. P. Soman "Diabetes detection using deep learning algorithms" ICT express, vol. 4, no. 4, pp.243-246, 2018

[11] M. Rahman, D. Islam, R. J. Mukti and I. Saha "A deep learning approach based on convolutional LSTM for detecting diabetes", Computational Biology and Chemistry, vol. 88, pp.107329, 2020.

[12] S. K. Das, P. Roy and A. K. Mishra "Deep Learning Techniques Dealing with Diabetes Mellitus: A Comprehensive Study", In Health Informatics: A Computational Perspective in Healthcare, pp. 295-323, 2021.

[13] T. Vijayakumar, R. Vinothkanna, and M. Duraipandian, "Fusion based feature extraction analysis of ECG signal interpretation–a systematic approach", Journal of Artificial Intelligence, vol. 3, no. 01, pp.1-16. 2021.

[14] J. I. Z. Chen, and P. Hengjinda, "Early Prediction of Coronary Artery Disease (CAD) by Machine Learning Method-A Comparative Study", Journal of Artificial Intelligence, vol. 3, no.01, pp.17-33, 2021.

[15] A. Sathesh "Computer Vision on IOT Based Patient Preference Management System", Journal of Trends in Computer Science and Smart Technology, vol. 2, no.2, pp.68-77, 2020.

[16] A. Sungheetha, "COVID-19 Risk Minimization Decision Making Strategy Using Data-Driven Model", Journal of Information Technology, vol. 3, no. 01, pp.57-66, 2021.

[17] A. Sharma, K. Guleria, and N. Goyal, "Prediction of Diabetes Disease Using Machine Learning Model", In International Conference on Communication, Computing and Electronics Systems, pp. 683-692, 2021.

[18] "Dataset," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

[19] P. Ghosh, A. Karim, S. T. Atik, S. Afrin, and M. Saifuzzaman, "Expert cancer model using supervised algorithms with a lasso selection

[20] P. Sharma and J. Singh, "Machine learning based effort estimation using standardization," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018.

[21] M. F. Foysal, M. S. Islam, A. Karim, and N. Neehal, "Shot-net: A convolutional neural network for classifying different cricket shots," Communications in Computer and Information Science, pp. 111–120, 2019.

[22] Z. Tasnim, S. Chakraborty, F. M. Shamrat, A. N. Chowdhury, H. A. Nuha, A. Karim, S. B. Zahir, and M. M. Billah, "Deep learning predictive model for colon cancer patient using CNN-based classification," International Journal of Advanced Computer Science and Applications, vol. 12, no. 8, 2021.

[23] K. R. Prilianti, T. H. P. Brotosudarmo, S. Anam, and A. Suryanto, "Performance comparison of the convolutional neural network optimizer for photosynthetic pigments prediction on plant digital image", In AIP Conference Proceedings, vol. 2084, no. 1, pp. 020020. 2019.

[24] U. Khair, H. Fahmi, S. A. Hakim, and R. Rahim, "Forecasting error calculation with mean absolute deviation and mean absolute percentage error," Journal of Physics: Conference Series, vol. 930, p. 012002, 2017.

[25] A. Islam Chowdhury, M. Munem Shahriar, A. Islam, E. Ahmed, A. Karim, and M. Rezwanul Islam, "An automated system in ATM booth using face encoding and emotion recognition process," 2020 2nd International Conference on Image Processing and Machine Vision, 2020.

[26] S. Kido, Y. Hirano, and N. Hashimoto, "Detection and classification of lung abnormalities by use of convolutional neural network (CNN) and regions with CNN features (R-CNN)," 2018 International Workshop on Advanced Image Technology (IWAIT), 2018.

[27] P. Ghosh, S. Azam, K. M. Hasib, A. Karim, M. Jonkman and A. Anwar, "A Performance Based Study on Deep Learning Algorithms in the Effective Prediction of Breast Cancer," 2021 International Joint Conference on Neural Networks (IJCNN), 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9534293.

[28] P. Ghosh, F. M. Javed Mehedi Shamrat, S. Shultana, S. Afrin, A. A. Anjum and A. A. Khan, "Optimization of Prediction Method of Chronic Kidney Disease Using Machine Learning Algorithm," 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2020, pp. 1-6, doi: 10.1109/iSAI-NLP51646.2020.9376787.

[29] S. Zobaed, F. Rabby, I. Hossain, E. Hossain, S. Hasan, A. Karim, and K. Md. Hasib, "Deepfakes: Detecting forged and synthetic media content using machine learning," Advanced Sciences and Technologies for Security Applications, pp. 177–201, 2021.