Proceedings of the 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT-2024)
IEEE Xplore Part Number: CFP24CV1-ART; ISBN: 979-8-3503-2753-3

Early Prediction of Diabetes and its Risk Factors based on ARIMA-ELMAN ANN Network

1. Dr. J. Senthil, Assistant Professor (Senior Grade), Department of Food Technology, Paavai Engineering College (Autonomous), Namakkal, Tamilnadu, India, senjana25@gmail.com

2. Shaik Akbar, Associate Professor, Department of CSE(AI & ML), Geethanjali College of Engineering and Technology, Hyderabad, India, shaikakbar.cse@gcet.edu.in

3. Akiladevi N, Assistant Professor, Department of Mathematics, Sri Eshwar College of Engineering, Kondampatti, Coimbatore, India, akiladevi83@gmail.com

4. S Praveena, Associate Professor, Dept of ECE, Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad, Telangana, India. spraveena_ece@mgit.ac.in

5. Ravindar K, School of Sciences, SR University, Warangal, Telangana, India. koppularavindar@gmail.com

6. Banupriya V, Assistant Professor, Department of Computer Science and Business Systems, M.Kumarasamy College of Engineering, Karur, India, banucs03@gmail.com

2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) | 979-8-3503-2753-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/IDCIOT59759.2024.10468045

***Abstract* –The prevalence of chronic diabetes is high. If diabetes could be predicted sooner, treatment results could be better.One such use of data mining techniques is in the field of disease prediction and early detection. To illustrate the potential value of these traits in predicting the presence or absence of diabetes, it is necessary to first explain the relationship between them. Technologies execute a wide variety of activities connected to diabetes, including clustering, association rule mining, and significant attribute selection determination. Priorities should be data preprocessing, feature selection, and model training. When it comes to data preprocessing, the cleansing and cleansing stage is in charge of dealing with errors, missing values, and inconsistencies. It takes advantage of the efficiency of the clustering algorithm and the reversing operation to exclude features with low F-scores. Prior to training ARIMA-ELMAN-ANN models, feature selection is necessary. The proposed approach clearly outperforms the two front-runners, ANN and ELMAN. Accuracy increased by 96.31 percent when the strategy was put into action.**

***Keywords*—Autoregressive Integrated Moving Average (ARIMA), Diabetes Mellitus (DM), Artificial Neural Network (ANN).**

I. INTRODUCTION

Medical experts typically use the term "diabetes mellitus" to refer to the disease. It's a metabolic disorder in which blood sugar levels remain consistently low. Diabetes affects a large population and can be divided into two types: Type 1 and Type 2. Children are disproportionately affected by type 1 diabetes, often known as insulin-dependent diabetes. Antibodies produced by the body assault the pancreas, leading it to stop generating insulin and ultimately wreak havoc on the body's internal organs, resulting in type 1 diabetes. Type 2 diabetes is also known as adult-onset diabetes or non-insulin-dependent diabetes. Even though type 2 diabetes is less severe than type 1, complications involving the eyes, kidneys, or nerves can prove fatal. According to statistics compiled by the WHO and released, the prevalence of diabetes among adults nearly doubled during that time period. The World Health Organization found that 3–16% of pregnant women have diabetes. Research into diabetes mellitus is a key priority in the medical sciences due to the significant social and economic effect of the disease.Therefore, machine learning and data mining tactics are of essential importance in diabetes mellitus when it comes to diagnosis, management, and other associated clinical administration features. The typical method of diagnosis involves patients to go to a clinic, consult with a doctor, and then wait a day or more before receiving their results. Plus, it's costly for them to have to induce their report of diagnosis every time. Type one polygenic condition is the form of diabetes in which the exocrine gland fails to produce the hormone hypoglycemic agent. It had previously been referred to as an endocrine dependent polygenic condition and an autoimmune disorder. Patients with this condition are among the minority, and it require the use of an artificial endocrine to supplement their own, either through injection or an endocrine pump. To refer to a collection of metabolic diseases in which abnormal hypoglycemic agent secretion and/or action is the fundamental cause of disease progression, the term "diabetes mellitus" (DM) is commonly used. Lack of hypoglycemic agents causes hyperglycemia, or high blood sugar, and slows down the metabolism of fat and protein. People with diabetes mellitus (DM) may experience an increase or decrease in their blood sugar levels. If this illness is not addressed, it could lead to serious complications or even death. People with DM often have compromised health, which can lead to the failure of a variety of vital organs. Of course, this calls for a prior diagnosis of diabetes. The DM health data have been utilized to efficiently evaluate a number of potential remedies to critical challenges. In a typical predictor, training record sets are mined for input, and the records are filtered using suitable pre-processing techniques. Classifiers are then used to make a prediction based on these inputs. Record sets are then tested using these trained models to confirm accuracy. In this proposed approach to describe a hybrid approach for predicting DM risk. The prevalence of diabetes is increasing rapidly in both urban and suburban settings nowadays. Early diabetes may be difficult to recognize due to the complexities of the procedure and other indicators, which may delay treatment. Recognizing the life-saving potential of early diabetes identification in supporting healthcare

practitioners inspired us to build a healthcare model. As a result, it is essential to develop a diabetes prediction system that can help doctors quickly and accurately identify people who are at higher risk for developing the disease. Machine learning techniques are crucial for predicting the start of many chronic diseases. Using machine learning on large health datasets has helped improve diagnostic accuracy while reducing the cost of problem diagnosis.

II. LITERATURE SURVEY

A chronic disease or disorder is one that persists over an extended period of time or has far-reaching effects. The impairment of quality of life is a major drawback of these conditions. One of the most devastating diseases, diabetes is a global epidemic. [1]This persistent disease is a major killer of working-age adults around the globe. Having a chronic illness comes at a financial cost. Governments and people alike spend a considerable sum on chronic diseases.[2]Despite its modest scope, study on biological data has, over time, opened the door to computational modelinganalytical frameworks based on statistics. Similarly, healthcare organizations are amassing a wealth of information. In machine learning, models are built to "learn" from the data they are given using targeted techniques, collect relevant data.In quantitative studies, diabetes diagnosis is seen as a challenging topic. The levels of several hematological indicators [3] were measured faults rendered it useless, therefore it wasn't really useful. These indications were employed to detect diabetes in investigations [4]. It has been hypothesized that just a few treatments, such as long-term ingestion of, can raise A1C. Medications, booze, and pain relievers when measured by electrophoresis, concentrations may appear to be dropping, even though they may be stable as determined by chromatographic analysis. It appears that the vast majority of studies that inflammation contributes to a healthy number of white blood cells in the body in spite of hypertension [5]linked with Insulin and Body Mass Index. Obesity, on the other hand, is not inextricably connected to having a potbelly.Diabetic mellitus (DM) is the official medical term for this disorder. Elevated blood sugar levels are a hallmark of metabolic illnesses like type 2 diabetes mellitus, type 1 diabetes, and pre diabetes. Long-term consequences of diabetes include cardiovascular disease, stroke, renal failure, heart attack, peripheral artery disease, and damage to the blood vessels and nerves [6]. From 122 million in 1980 to 422 million, the global prevalence of diabetes has skyrocketed. In the year 2040, there will be approximately 642 million people [7]. In addition, around 1.6 million deaths occurred all over the world in 2014 because of diabetes [8]. As a result, this is a figure that seriously worries us. As the prevalence of diabetes rises, so does the number of people losing their lives to the disease. Type I diabetes, often known as juvenile onset diabetes, is the most severe form of diabetes [9]. Other types of diabetes include type II and pregnancy-related diabetes. Most persons with type 1 diabetes are under the age of 30. A number of symptoms, including polyuria, thirst, frequent hunger, weight loss, vision issues, and fatigue, can indicate type 1 diabetes. Type 2 diabetes is associated with being overweight, having high blood pressure, having abnormal cholesterol levels, and having atherosclerosis [10]. Thirdly, there's gestational diabetes. Gestational diabetes does affect pregnant women. Most medical data are nonlinear, no normal, correlation organized, and complex, making diabetic data analysis a difficult job. When it comes to medical imaging, ML-based systems have emerged as the frontrunners. This is true across a wide range of diagnostic areas, from stroke to coronary artery disease to cancer.Diabetes is now officially an international health crisis. When hyperglycemia persists untreated for a long time, it can irreversibly impair various physiological systems. Early detection and treatment are key to preventing or delaying diabetes and its consequences. However, due to budgetary constraints and a lack of health literacy, only half of people with diabetes are now diagnosed [11]. A disease risk assessment model built with physical examination data could provide clinical guidance and early, widespread screening in light of the exponential expansion in physical examination data and the rapid development of artificial intelligence [12]. The 2019 coronavirus (COVID-19) outbreak has been linked to preexisting conditions including diabetes, according to researchers. In addition, knowing an individual's risk for high blood sugar can help researchers devise new methods of warding off and treating cardiovascular disease [13]. Numerous studies have focused on the creation of various diabetes detection models. We employed multivariate logistic regression [14] to enhance the AUC on the basis of shared characteristics. Some of the quirks of laboratory data render these models useless, resulting in poor performance. The feature selection method of developing a prediction model was first presented in another research work. The dataset utilized in the research was obtained from Kaggle, a machine learning repository, and contained certain inaccuracies when compared to the real world. Predicting diabetes in the Chinese population was the focus of [15] model; however, no system or diabetes risk assessment was included in the study. Due to a lack of data from rigorous laboratory testing and insufficient training data, the majority of these models also fell short of expectations.[16]Most people nowadays are too preoccupied with work and other commitments to give much thought to their health or to taking preventative measures. As a result, we may be more likely to develop lifestyle-related diseases like diabetes mellitus. It's possible that this disease could be the deadliest one ever recorded if it's not caught in time [17]. Glucose in the blood, which comes mainly from the food you eat, is the body's primary source of energy. [18]The production of insulin makes the pancreas an important organ. Insulin is a hormone whose major role is to maintain constant glucose (blood sugar) levels in the blood. Glucose is a vital metabolic fuel that can be obtained from carbohydrates in the food we eat. When T2DM is well managed, complications from the disease are less likely to occur [19]. Type 2 diabetes mellitus (T2DM) is an essential metabolic condition that has both micro vascular and macro vascular effects. At present, glycated hemoglobin (HbA1c) is the most reliable indicator of diabetes control. However, HbA1c's inability to capture daily variations in

glucose regulation remains a serious constraint in the monitoring of diabetic patients [20], despite the fact that it will continue to play a key role in assessing the level of diabetic control for the foreseeable future. Uric acid is a waste product of purine metabolism, and research has connected it to metabolic dysfunction and kidney disease. Uric acid levels are often greater in persons with type 2 diabetes [21]. Hypothesized to be an indicator of metabolic syndrome, decreased HDL cholesterol has also been associated to a deterioration in metabolism. The uric acid to HDL cholesterol ratio (UHR) combines these two measurements to create a more accurate indicator of metabolic deterioration.

III. PROPOSED SYSTEM

Worldwide, millions of individuals live with diabetes, a chronic disease that can last a patient a lifetime. Diabetes impacts individuals of all age groups. Innovation in technology is a novel approach to diabetes prediction that improves both efficiency and accuracy. The majority of research has been on diabetes prediction, with the Pima Indian dataset seeing the most extensive use. The authors of this proposed to build a framework to precisely assess the likelihood of people having diabetes.

A. Data Preprocessing:

Data gathered from real-world issues isn't necessarily complete or accurate in many domains, healthcare included. Data cleansing is the process of removing errors and inconsistencies from a dataset. Over the years, many solutions to this issue have emerged. Prior to choosing a strategy, many factors should be taken into account. In order to deal with incorrect data correctly, follow these steps:

1) Discarding the Missing Values:

The most usual approach to discard the MVs but this approach is not so practical because if the train data have a large number of missing values then the produced result must biased. If the dataset has a small number of missing values, then we must assure that analysis on the remaining part will not produce the inference bias. The deletion of the MVs can be done in following ways:

a) Listwise Deletion:

A comprehensive case analysis is performed to do this, and cases where more than one value is missing are eliminated. But in which aren't many blanks in the data, this approach shines[22]. It works wonderfully with datasets that contain the rare MCAR missing pattern.

b) Dropping Attribute Completely:

It may be more prudent to eliminate the property completely in instances where there are more than 62% missing observations and the feature does not appear to possess any statistical significance; nevertheless, this is an atypical scenario. Due to their high significance, attributes with missing values should occasionally be maintained. A lack of confidence in the usefulness of missing cases or qualities is no excuse to disregard them. Consequently, specific probabilistic techniques and imputation procedures are usually prioritized when dealing with MVs.

c) Pairwise Deletion:

A list-wise deletion with a smaller margin of error is the target of this method. An MV-containing attribute is deleted from the data set if it is not being utilized as a case for any other attribute. Although it increases the power of analysis, it complicates things (such as defining standard error).

B. Feature Selection:

Subsequently, the proposed Framework proposes using the PID dataset to choose distinctive features. Having moved on to a straightforward method known as the F-Score, which assesses the differentiating factor between two classes using real values, after it removed characteristics from the dataset that were either missing values or had a missing data rate more than 6%. The wrapper component re-studies the characteristics with reasonably high F-scores to further explore their contributions to accurate clustering. People are then considered for the informational feature based on their potential. It started by removing the features with the lowest F-scores and working their way backwards to see how effective the grouping was[23]. In the second stage of the suggested framework, let will apply the following equation to get the F-Score values for each PID feature:

$$F(g) \equiv \frac{\left(q_g^{(+)} - q_g\right)^2 + \left(q_g^{(-)} - q_g\right)^2}{\frac{1}{o_+ - 1}\sum_{v=1}^{o_+}\left(q_{v,g}^{+} - q_g^{+}\right)^2 + \frac{1}{o_- - 1}\sum_{v=1}^{o_+}\left(q_{v,g}^{-} - q_g^{-}\right)^2} \quad (1)$$

In the initial data pre-processing step, after removing any features or feature instances with missing data, it can use Eq. (1) to get the F-Score value for each feature that choose. $q_g^{(+)}, q_g^{(-)}, q_g$ Each feature is being categorized into three groups: positive, negative, and all data. The average value of all instances is represented by this. Whereas $q_{v,g}^{+}, q_{v,g}^{-}$ stand for the value in every instance of the feature.In order to do discriminative analysis on each PID feature, can apply Eq. (1) to find their F-Score. A feature called "superfluous" can be removed from an F-Score in order to increase the possible feature usage.

C. Model Training:

1) ARIMA:

Imagine a time series that includes an observation s_p and a mean of ρ. The acronym ARMA describes a mixed autoregressive moving average model, which is

$$\phi(W)\bar{s}_p = \theta(W)r_p \quad (2)$$

where W represents the random error of the time series r_p time p and $(\bar{s}_p = s_p - \rho)$ is the equation for the backward shift operator. The parameters of the model are $\phi_g\,(g = 1,2,\dots,t)$ and $\phi_h\,(h = 1,2,\dots,u)$, and the orders of the model are generally denoted by the numbers t and u. Both $\phi(W) = 1 - \sum_{g=1}^{t}\phi_g w^g$ and $\theta(W) = 1 - \sum_{h=1}^{x}\theta_h w^h$ are polynomials of degree t and x, correspondingly. The arbitrary errors are likewise assumed to adhere to a normal distribution with a constant variance of ω^2 and a mean of zero.If the roots of $\phi(W) = 0$ are located outside

the unit circle, the ARMA process is stationary. Their explosive non-stationary behavior is only observed when it occurs inside the unit circle. If the autoregressive operator $\phi(W)$ is stationary, then the ARIMA process can be expressed as using

$$\phi(W)(1-W)^f \bar{s}_p = \theta(W) r_p \quad (3)$$

where the order of difference is frequently denoted by the integer f. When the operators $\nabla = 1 - W$ and $\nabla^f \bar{s}_p = \nabla^f r_p$ are introduced, the equation given above becomes

$$\phi(W)\nabla^f s_p = \theta(W) r_p \quad (4)$$

ARIMA(t, f, x) is the expression of the ARIMA model when different values of t, f and x is considered. According to ARIMA, the future value of a variable is believed to be linearly related to a number of prior observations plus random mistakes.Once the function of the ARIMA model has been specified, the next step is to estimate its parameters. The goal in making parameter estimates is to lower the total measure of error. A least squares estimate approach is typically used for this. One last step before finishing a model is to make sure it's adequate through diagnostic testing. Its main purpose is to check if the model's assumptions about the errors are true. With the help of several diagnostic tools and residual plots, one can may check how well the model that is being considered fits the historical data. If the existing model is inadequate, it is recommended to find a new tentative model and then repeat parameter estimates and model verification.

2) ELMAN:

In order for an Elman neural network to function, the activation levels of the input units are set up in a specific pattern. The activation value for each hidden unit is obtained by multiplying the input and context activation values with the weight value. Next, further add the bias of the hidden unit to the total, and then simply compress the result with a function d. Then, the output of the hidden unit is the value that comes out of it. Elman neural networks use a logistic d squashing function. Using the same method, then can calculate the activations of the output units[24]. In this instance may see a single time step. Each hidden unit's activation is first reproduced in its matching context unit with a one-to-one basis and fixed weights of 1 before going to the next time step. By linking each hidden unit to itself, this method limits access more so than the free-form recurrent connections.What follows is a hypothetical situation in which n is the number of input units, l is the number of hidden units, and m is the number of output units. As an example, the network can accept the input $q_g (g = 1,2,\dots,o)$ and produce an output $u_v (v = 1,2,\dots,c)$ .The hidden and output biases are represented by w_h and w_v, respectively, and the functions $d(.)$ and $i(.)$ are used to calculate them. A mathematical diagram of an Elman neural network is shown below.The hidden module's output:

$$u'_h(p) = d\left(\sum_{g=1}^{o} b_{hg} q_g + \sum_{a=1}^{e} b_{ha} u'_a(p-1) + w_h\right) \quad (5)$$

The outcome of the unit's output:

$$y_v(p) = i\left(\sum_{h=1}^{e} b_{hv} u'_h(p) + w_v\right) \quad (6)$$

The classic neural network cannot handle patterns that evolve over time. However, there is some temporal volatility in the inputs to the neural network, which are historical economic series data. Because of this, it should choose a network that can perform long-term data analysis. It aims to use an Elman recurrent neural network to forecast economic time series residual series.

3) ANN:

The goal of ANNs, or artificial neural networks, is to mimic the brain's pattern recognition and learning capabilities. Although other models of artificial neural networks (ANNs) have been proposed, the feedforward neural network (FNN) has had the greatest uptake. Training and testing are the two main phases of a FNN's learning process. During training, a training pair is defined as (q, u), where q is the vector of the independent variable and u is the vector of the dependent variable. After that, the configuration of the input and output layers is adjusted to match this pair. The objective is to construct the following equation:

$$u = d(q) = d_r(q) = d(q; r), r \in R \quad (7)$$

where $r \in R$ family of explicitly parameterized models is used to define the function $d = d_r$. A standard FNN training run involves minimizing the error for each training pair in order to fine-tune the connection-weight matrices. The training procedure is ended and the final connection weight matrices are kept for testing after all training pairings attain an acceptable average squared error.Time series forecasting makes extensive use of FNNs due to the fact that FNNs' nonlinear modeling capabilities effectively capture time series' nonlinear features. The end result of using FNN for time series forecasting is best shown as

$$u_p = r_0 + \sum_{h=1}^{o} b_h d\left(r_h + \sum_{g=1}^{c} b_{gh} u_{p-g}\right) + \psi_p \quad (8)$$

Here, the variables $r_h (h = 0,1,2,\dots,o)$ and $b_{gh} (g = 1,2,\dots,c; h = 1,2,\dots,o)$ stand for the bias on the h-th unit, $d(.)$ is the hidden layer's transfer function, m is the number of input nodes, and o is the number of hidden nodes.More specifically, Equation (7) shows that the FNN model uses a nonlinear function to convert the observed values from the past $u_{p-1}, u_{p-2}, \dots, u_{p-c}$ into the future value u.

$$u_p = \tau(u_{p-1}, u_{p-2}, \dots, u_{p-c}, \kappa) + \psi_p \quad (9)$$

ψ is a function that is determined by the network structure and connection weights, and κ is the parameter vector. So, the FNN model is similar to a nonlinear autoregressive model in several ways.

IV. Result and Discussion

A high blood glucose level and several other metabolic problems result from cells not responding properly to insulin or the body not producing enough insulin in a diabetic person. Organs such as the kidneys, eyes, heart, nerves, and veins are particularly susceptible to the harm, damage, and eventual failure that can be brought about by the persistent hyperglycemia that is a hallmark of diabetes. In order to get findings that are similar to clinical outcomes, this proposed aims to utilize significant features, construct a machine learning prediction method, and determine the best classifier. Finding the specific traits that aren't good predictors of Diabetes Miletus is the recommended strategy for early detection using predictive analytics.

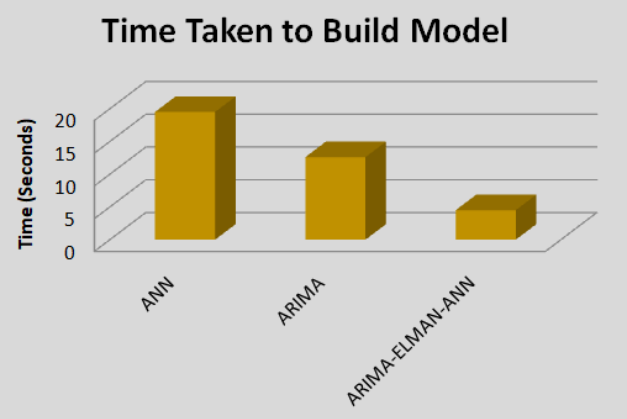

Fig. 1. Time Taken to Build Classification Model

Figure 1 shows that the time it took to construct the classification model was relatively short, with the longest being 19 seconds for ANN. However, when ARIMA began using feature selection for its classification model, the time it took to construct the model increased dramatically, reaching 12 seconds. Models that were suggested had the quickest time of 4.2 seconds.

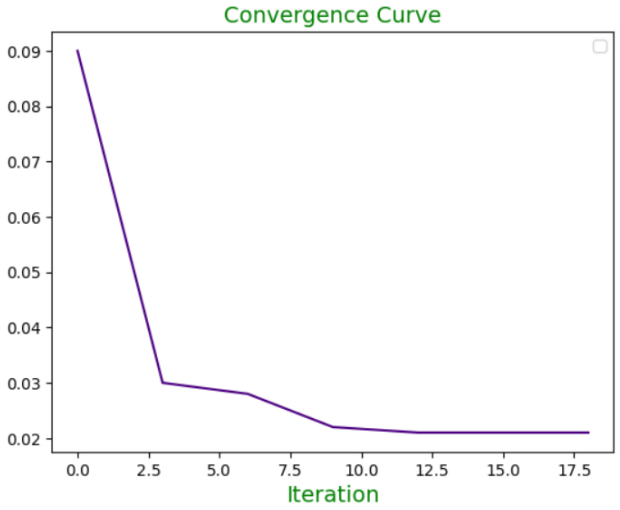

Fig. 2. Convergence Curve of the Fitness Function

Figure 2 shows the 18 iterations in relation to the fitness function value. An increasing iteration count will result in a smaller minimization function. With 18 iterations to have achieved an ideal fit value of 0.021.

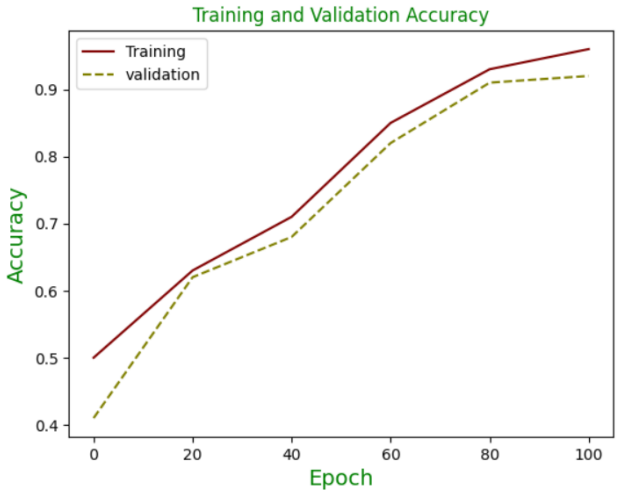

Fig. 3. Training and Validation Accuracy

After 100 training epochs, the ARIMA-ELMAN-ANN approach reaches an accuracy of approximately 96.43%, as shown in Figure 3. Then, to get the results shown in Figure 3, this subset of features is trained on the ARIMA-ELMAN-ANN using 100 epochs.

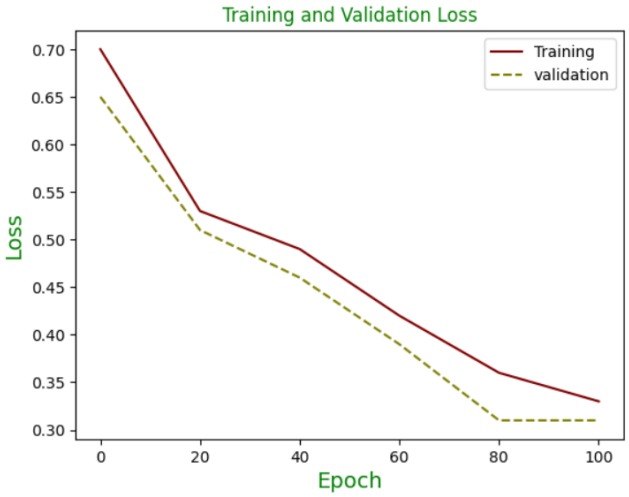

Fig. 4. Training and Validation Loss

On average, every 20 epochs, a new loss performance is added into the GWO research to represent a test error (figure 4). The model would be overfit if they continue along the current upward trend. Stopping at this point is premature.

V. Conclusion

When insulin production drops below normal levels, a dangerous condition called diabetes mellitus sets place. Insulin is responsible for maintaining normal blood glucose levels. Delays in diagnosis can cause harm to several bodily systems, such as the nerves, eyes, and kidneys. New technological developments have piqued people's interest in personalized healthcare. A fast growing subfield of predictive analytics, machine learning is finding extensive use in healthcare for the purpose of

early disease identification and symptom prediction. The main objective of this proposed is to create a model that, using machine learning classification algorithms and data related to the condition, can forecast when diabetes will start. The cleaning and cleansing stage of data preprocessing is responsible for handling mistakes, missing values, and discrepancies. The goal is to exclude features with low F-scores by utilizing the reversing technique and the efficient clustering approach. The first stage in training ARIMA-ELMAN-ANN models is feature selection. Compared to the other two methods, ARIMA-ELMAN-ANN performs much better(around 96.31%).

REFERENCES

[1] A. Chaturvedi, L. Mohapatra, A. Jain, S. Emn, D. Suganthi, and R. V. Srinivas, "An Innovative Approach of Early Diabetes Prediction using Combined Approach of DC based Bidirectional GRU and CNN," in *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Jul. 2023, pp. 947–952. doi: 10.1109/ICESC57686.2023.10193133.

[2] J. S. Skyler *et al.*, "Differentiation of diabetes by pathophysiology, natural history, and prognosis," *Diabetes*, vol. 66, no. 2, pp. 241–255, 2017, doi: 10.2337/db16-0806.

[3] B. Dorcely *et al.*, "Novel biomarkers for prediabetes, diabetes, and associated complications," *Diabetes, Metab. Syndr. Obes.*, vol. 10, pp. 345–361, 2017, doi: 10.2147/DMSO.S100074.

[4] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. N. Davis, "DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values," *IEEE Access*, vol. 7, pp. 102232–102238, 2019, doi: 10.1109/ACCESS.2019.2929866.

[5] H. N. Merad-boudia, M. Dali-Sahi, Y. Kachekouche, and N. Dennouni-Medjati, "Hematologic disorders during essential hypertension," *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 13, no. 2, pp. 1575–1579, 2019, doi: 10.1016/j.dsx.2019.03.011.

[6] A. Krasteva, V. Panov, A. Krasteva, A. Kisselova, and Z. Krastev, "Oral cavity and systemic diseases - Diabetes mellitus," *Biotechnol. Biotechnol. Equip.*, vol. 25, no. 1, pp. 2183–2186, 2011, doi: 10.5504/bbeq.2011.0022.

[7] P. Zimmet, K. G. Alberti, D. J. Magliano, and P. H. Bennett, "Diabetes mellitus statistics on prevalence and mortality: Facts and fallacies," *Nat. Rev. Endocrinol.*, vol. 12, no. 10, pp. 616–622, 2016, doi: 10.1038/nrendo.2016.105.

[8] N. Saravanan, S. Venkatalakshmi, and C. Bharath, "Assessment of knowledge related to diabetes mellitus among patients attending a dental college in Salem city-A cross sectional study," *Brazilian Dent. Sci.*, vol. 20, no. 3, pp. 93–100, 2017, doi: 10.14295/bds.2017.v20i3.1437.

[9] G. Danaei *et al.*, "National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: Systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2·7 million participants," *Lancet*, vol. 378, no. 9785, pp. 31–40, 2011, doi: 10.1016/S0140-6736(11)60679-X.

[10] G. Robertson, E. D. Lehmann, W. Sandham, and D. Hamilton, "Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: A proof-of-concept pilot study," *J. Electr. Comput. Eng.*, vol. 2011, 2011, doi: 10.1155/2011/681786.

[11] R. L. Thomas, S. Halim, S. Gurudas, S. Sivaprasad, and D. R. Owens, "IDF Diabetes Atlas: A review of studies utilising retinal photography on the global prevalence of diabetes related retinopathy between 2015 and 2018," *Diabetes Res. Clin. Pract.*, vol. 157, p. 107840, 2019, doi: 10.1016/j.diabres.2019.107840.

[12] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.

[13] G. P. Fadini, M. L. Morieri, E. Longato, and A. Avogaro, "Prevalence and impact of diabetes among people infected with SARS-CoV-2," *J. Endocrinol. Invest.*, vol. 43, no. 6, pp. 867–869, 2020, doi: 10.1007/s40618-020-01236-2.

[14] W. Bao *et al.*, "Predicting risk of type 2 diabetes mellitus with genetic risk models on the basis of established genome-wide association markers: A systematic review," *Am. J. Epidemiol.*, vol. 178, no. 8, pp. 1197–1207, 2013, doi: 10.1093/aje/kwt123.

[15] X. Zhou *et al.*, "Nonlaboratory-based risk assessment algorithm for undiagnosed type 2 diabetes developed on a nation-wide diabetes survey," *Diabetes Care*, vol. 36, no. 12, pp. 3944–3952, 2013, doi: 10.2337/dc13-0593.

[16] R. M. Khalil and A. Al-Jumaily, "Machine learning based prediction of depression among type 2 diabetic patients," *Proc. 2017 12th Int. Conf. Intell. Syst. Knowl. Eng. ISKE 2017*, vol. 2018-Janua, no. October, pp. 1–5, 2017, doi: 10.1109/ISKE.2017.8258766.

[17] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[18] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0175-6.

[19] G. Aktas *et al.*, "Mean Platelet Volume (MPV) as an inflammatory marker in type 2 diabetes mellitus and obesity," *Bali Med. J.*, vol. 7, no. 3, pp. 650–653, 2018, doi: 10.15562/bmj.v7i3.806.

[20] T. D. Control, "Modern-Day Clinical Course of Type 1 Diabetes Mellitus After 30 Years' Duration," *Arch. Intern. Med.*, vol. 169, no. 14, p. 1307, 2009, doi: 10.1001/archinternmed.2009.193.

[21] M. Z. Kocak, G. Aktas, E. Erkus, I. Sincer, B. Atak, and T. Duman, "Serum uric acid to HDL-cholesterol ratio is a strong predictor of metabolic syndrome in type 2 diabetes mellitus," *Rev. Assoc. Med. Bras.*, vol. 65, no. 1, pp. 9–15, 2019, doi: 10.1590/1806-9282.65.1.9.

[22] P. Misra and A. S. Yadav, "Impact of Preprocessing Methods on Healthcare Predictions," *SSRN Electron. J.*, no. January, 2019, doi: 10.2139/ssrn.3349586.

[23] R. B. Lukmanto, Suharjito, A. Nugroho, and H. Akbar, "Early detection of diabetes mellitus using feature selection and fuzzy support vector machine," *Procedia Comput. Sci.*, vol. 157, pp. 46–54, 2019, doi: 10.1016/j.procs.2019.08.140.

[24] Y. Xiao, J. Xiao, and S. Wang, "A hybrid model for time series forecasting," *Hum. Syst. Manag.*, vol. 31, no. 2, pp. 133–143, 2012, doi: 10.3233/HSM-2012-0763.