

Hybrid Machine Learning Algorithms Health Database for Diabetes Prediction using MLP and SVM

Kumar Gaurav¹ and Dr.Kakoli Banerjee²

¹⁻²JSS Academy of Technical Education/CSE, Noida, India

E mail: believegaurav22@gmail.com, Kakoli.banerjee@jssaten.ac.in

Abstract— The presence of persistently increased blood glucose levels in diabetic individuals serves as a diagnostic indicator of the condition. Diabetes mellitus type 2 is characterised by the inability of cells inside the body to make use of insulin that has been created. This is the hypothesis put out by medical professionals who study type 2 diabetes. The human organ known as the pancreas, which is responsible for producing insulin, may become unable to do its job properly, which can lead to the development of diabetes mellitus type 1. There has been a huge increase in the number of extremely sensitive health data as a direct result of the improvements that have been made in biotechnology as well as the strengthening of the public healthcare system. Intelligent data analysis technologies have revealed a surprising number of patterns. These technologies were built with the intention of preventing a variety of severe diseases. Diabetes mellitus puts a person at a much greater risk of developing cardiovascular disease, kidney disease, and nerve damage, all of which may significantly shorten a person's life expectancy. In this study, a hybrid learning algorithm is formed by combining the MLP multilayer perceptron method with the SVM support vector machine algorithm. This results in improved accuracy in diabetes identification.

Index Terms— MLP, SVM, ML, Diabetes, Prediction

I. INTRODUCTION

Deep learning is a subfield of machine learning that falls under the umbrella of the field of artificial intelligence (AI). This topic focuses on the development of automated systems that are capable of recognising patterns in data sets for the aim of generating predictions and categorising fresh data [7]. Machine learning is an umbrella word that refers to a collection of methodologies for the development of models. These methodologies take inspiration from the disciplines of statistics, data mining, and optimization.

The purpose of providing a meaningful representation of the data-driven insights that are automatically created is the mission of representational learning [7] in the field of machine learning. We refer to the process of working with a multi-level representation if we are dealing with a complicated representation or whenever it is necessary to use deep learning approaches.

It will be simple to read and understand if it is presented in the form of a spreadsheet containing tabular data. The final version of the dataset can have as little as 100 cases or as many as 500 individuals altogether. A comparison of the improvement with the findings produced via the use of other methodologies, such as support vector machines, logistic regression, or the nave methodology, is another possibility. Depending on the procedures that are carried out, the diagnostic test will either provide a positive or negative result in order to identify or diagnose

the condition. It is feasible that the accuracy of the systems that are now in place might be significantly improved by the use of MLP and hybrid learning approaches.

It is possible that the accuracy of treatment plans and the results of medical care might be improved using machine learning technologies since these technologies depend on algorithms and algorithmic techniques. One of the numerous applications of deep learning in radiology and medical imaging that is seeing tremendous growth is the simulation of the functions performed by the brain. This subfield of machine learning is considered to be rather advanced. Deep learning techniques, which enable neural networks to learn from data on their own without being supervised, are currently being used in the process of locating, detecting, and evaluating malignant tumours in photographs. This approach was previously used. The application of machine learning algorithm specifications and improved cloud infrastructures have made it possible for rising processing rates to enable the identification of anomalies in photographs that would otherwise go unnoticed by a human doctor. These anomalies would otherwise go unnoticed. Learning machines will have a significant influence on the medical industry, which is undergoing fast change and will continue to do so.

II. LITERATURE REVIEW

According to Moreb et al. [1] The SEMLHI framework is available to academics and developers, who may use it to plan activities inside the framework using a SEMLHI methodology. The general public may use this authorization. The whole architecture is made up of four separate parts: software, artificial learning, artificial learning algorithms, and fitness informatics data. These modules enable software developers and researchers to conduct an engineering analysis of fitness informatics software, giving them a new road map for creating fitness applications that incorporate both system functions and software implementations. In the subject of fitness informatics, this represents a significant advancement. Beam et al. [2] People are changing their lives to make room for the advancements that are being made as a consequence of these two sectors as a result of the impact that big data and machine learning are having on nearly every aspect of modern life. This directly contributes to the modern world being more complex. Due to the history of users' search inquiries, both Netflix and Google are aware of the films that users select to view on their own platforms. Additionally, Google and Netflix are aware of the material that individuals are interested in knowing more about but cannot access.

III. PROBLEM STATEMENT

The ability to gather, share, and disseminate data has grown essential as digital disruption affects every industry, including healthcare. Massive amounts of data create a number of challenges that machine learning, big data, and artificial intelligence (AI) may help solve. Machine learning may also help healthcare organisations lower costs, better manage operations, and satisfy growing medical demands. Machine learning innovation at the point of treatment may help physicians identify and treat sickness more quickly, accurately, and with more individualised care. An examination of machine learning in healthcare reveals how technology developments could lead to more efficient, all-encompassing treatment approaches that might perhaps improve patient outcomes.

IV. PROPOSED METHODOLOGY

An effective way to increase accuracy is to use an extreme machine learning technique on health informatics prediction. Excessive learning machine (ELM) is a unique feedforward neural network learning technique that uses a single hidden layer and a large amount of input data. ELM is based on the empirical risk minimization hypothesis, which, unlike other approaches, can be learned in a single iteration. This strategy avoids many rounds and local reduction. A flowchart of the procedure outlined below is shown in Figure 1. As a result of issues like a lack of suitable hospitals or beds to accommodate an increasing number of patients, the author suggests combining Software Engineering with Machine Learning methods in the reference base paper [1] to improve disease prediction in healthcare systems and decrease the time required to anticipate disease.

V. IMPLEMENTATION

It's possible that a technique that combines MLP and hybrid learning with SVC might be helpful for making predictions in health informatics. To our good fortune, health informatics provide a mechanism to achieve this objective. The so-called "excessive learning machine" is an innovative method to the process of learning that makes use of neural networks (MLP). Since it has just one hidden layer, it is impossible for it to operate without a significant amount of training data. The MLP allows for the empirical risk minimization hypothesis to be learnt

with only one iteration, which is a significant time savings compared to the approaches that came before it. In light of the fact that this is the situation, the empirical risk minimization hypothesis is of great assistance. Flow diagram is shown in fig. 1.

In addition, the author of the foundational research [1] suggests that we reduce the amount of time required to anticipate illness owing to issues. This recommendation can be found in the paper. The author is of the opinion that the challenge of illness prediction would be able to be tackled in a manner that is both more expedient and fruitful if one were to combine the capabilities of software with those of machine learning. The writer of this book puts out a hypothesis that may be recalled by its initials.

Researchers need first and foremost to have access to relevant health informatics data in order to be able to develop trustworthy Machine Learning models for sickness prediction. In spite of this, missing data, null values, and other types of data that are not numeric are all too frequent in datasets, which may cause a loss of accuracy in machine learning predictions. Using PREPROCESSING, the author first cleans up the healthcare data by eliminating any values that are missing or null and then uses Python to transform the textual input into numerical form. Because of this, the author is able to come up with a solution that can really be implemented. It is possible to remove unnecessary fields or features from a dataset by using the PREPRO SKLEARN PCA method, which is classified as a dimensionality reduction technique. PCA improves the dataset in number of useful datasets.

For the duration of the class, the author makes use of several different machine learning strategies, some of which include This technique will automatically self-train itself using public datasets in order to produce a model. The application of the trained model to newly collected test data for the purpose of predicting is the next phase. You can educate a computer to learn and anticipate events on its own using the methods that we have been covering. The use of a machine algorithm model enables one to arrive at this prediction.

4) Software: In order for developers to guarantee the dependability of the aforementioned modules, it is necessary for them to make use of software quality assurance tools, software verification procedures, and UNIT TESTING. Using methods like as classification, clustering, and regression when dealing with datasets of varying sizes is one of the suggestions made by the author. The author utilised the PIMA dataset.

The MLP was used to the case study that included the acquisition of data linked to health. A innovative method that was created in the 1990s and given the name "multi-layer Perceptron" was developed with the goal of improving the instruction of neural networks by employing a single hidden layer (MLP). This approach sidesteps the issues of slow training speed and overfitting that afflict the conventional neural network learning methodology, which is the normal neural network learning method. This method is also known as the standard neural network learning method. A Multi-layer Perceptron has its own set of settings for each hidden node, and these parameters are learnt via a process of trial and error. They have a wide range of applications in the real world, some of which include, but are not limited to, regression analysis, classification, sparse approximation, clustering, feature learning, and compression. MLPs, which stands for multi-layered perceptrons, are a kind of neural network that is often used in deep learning.

Multiple-Layer Perceptrons (MLP) are versatile computational models that may be used to a wide variety of problems, including clustering, logical regression, and classification. The MLP uses the same feedforward neural network design as "generalised" single-hidden-layer networks as its foundation. In MLP, the creation of the parameters for hidden nodes requires the use of a random number generator. The MLP problem may be considered an optimization problem, using a formulation that is comparable to the optimization question for the support vector machine, if one were to judge it by the norms of classical optimization theory (SVM). On the other hand, SVM often produces less desirable results in comparison to MLP. Figure 3 presents a mathematical representation of the situation.

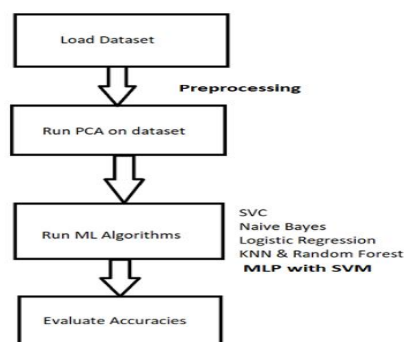


Figure 1. Flow Chart

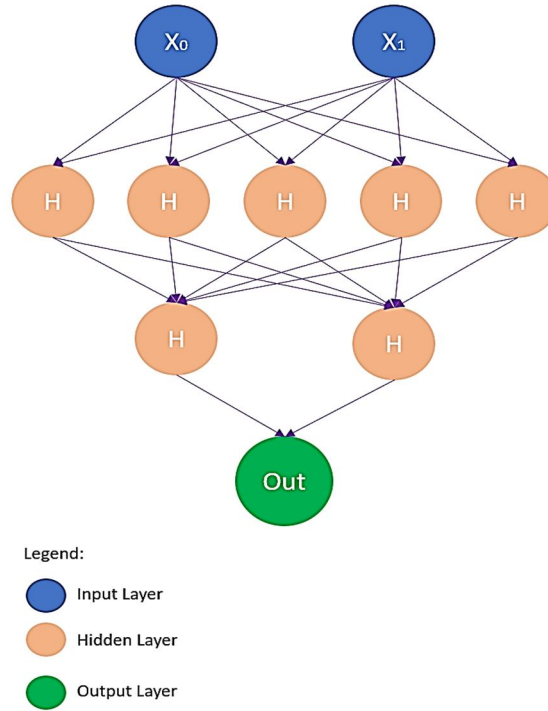


Figure 2. MLP Layers

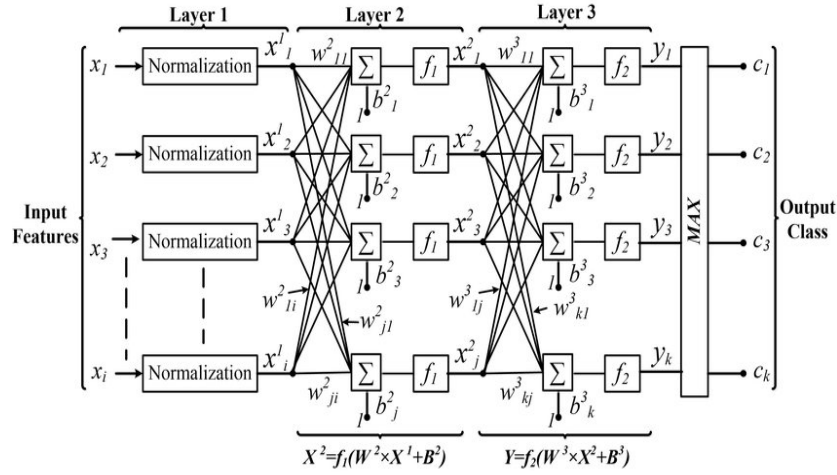


Figure 3. MLP Layers Mathematical Model

VI. RESULTS

The preprocessing stage is shown on the screen that follows. After that, principal component analysis (PCA) is performed on the data at this point in the process. By the use of a statistical method known as principal component analysis and the utilisation of an orthogonal transformation, correlated variables are possible to be converted into independent ones (PCA). In order to accomplish this objective, a fresh set of variable collections is developed and used. Principal Component Analysis is what "PCA" stands for when abbreviated. For exploratory data analysis and visualisation in machine learning, the principal component analysis (PCA) technique is the method of choice, and the principal component analysis (PCA) instrument is the instrument of choice in data science.

The titles of each of the various columns are indicated at the very top of the chart, where they may be easily seen. The machine learning algorithm won't be trained using columns with negative values. The results of the preprocessing that was performed on the data obtained from the various tests and exercises can be seen in the table that can be seen below.



Figure 4. PCA

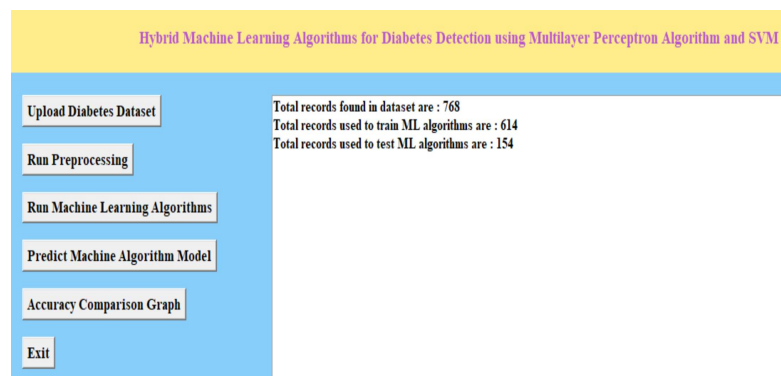


Figure 5. Record Mapping

After pre-processing and PCA, there were still 768 records visible on the screen. The application then used 1,614 of those data to instruct its machine learning algorithms and build models. When the model had been trained, it was evaluated using an additional 154 data points to see how well it could predict the values of those points. The next stage is to make use of the machine learning algorithms that were discussed in the subject on in-depth implementation that was covered the semester before. When the data needed for assessment and teaching have been acquired, this step is taken. Look at the results in the right-hand column; the calculator has already done the work for you.

At the time when these programmes are active, they carry out the following:

The KNN (K-Nearest Neighbor) technique was one of the first and most basic machine learning algorithms. The strategy is based on the principle of directed learning. K-Nearest Neighbors works by first making the assumption that the data of a new case are the same as that of the cases that came before it. This is how it determines which category the new case belongs in. The K-Nearest Neighbor (K-NN) approach is used in this situation to maintain track of all the information that has been obtained and to categorise newly acquired information according to the degree to which it is comparable to the material that has previously been stored. Hence, in the event that fresh data is generated, it is possible to rapidly classify it by using the K-NN algorithm along with an uncomplicated algorithmic technique.

The Bayes' Theorem and the assumption that predictors are not coupled provide the basis for the categorization of data using the Naive Bayes method. Naive Bayes classifiers are too simple because they operate on the assumption. The use of ensemble learning to solve difficult problems necessitates the cooperation of several independent classifiers.

The estimation of the occurrence probability of a target variable is a common application of the statistical classification method known as logistic regression, which is based on supervised learning. Before one can make predictions about a categorical dependent variable, one must first amass a collection of independent variables. Logistic regression is the technique of choice for generating predictions when the dependent variable is of the categorical form since it is easier to interpret the results. In order for the output to have any value whatsoever, it is

necessary for it to be discrete or categorised. Probabilities ranging from 0 to 1 are provided in lieu of concrete numbers wherever possible. These are sent irrespective matter the value of the answer that was provided (yes/no/zero/one/true/false, etc.).

Information may be categorised using a technique called nonparametric clustering, which does not presuppose anything about the data it is examining or the probable sizes of the clusters it creates. In this context, we are referring to a particular flavour of SVC known as linear SVC. In order to make high-dimensional data more comprehensible in advance of its use, preprocessing methods may be applied to it. Some examples of these methods are principal component analysis. Any algorithm that uses the MLP or hybrid learning algorithm will be affected by the changes brought about by this update.

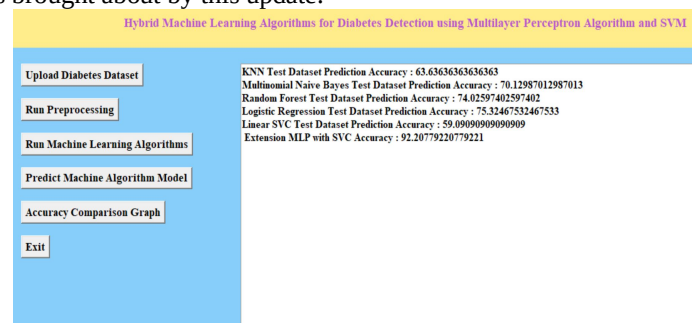


Figure 6. MLP Output

The previous screen displayed the maximum accuracy of the MLP Output, indicating that the machine learning technique had reached its maximum potential in terms of accuracy.

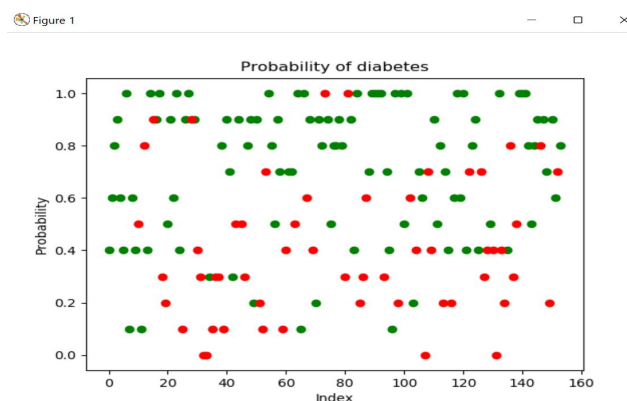


Figure 7. Chart for Probability

You can see your diabetes risk on the screen up top, with red being a good prognosis and green a negative one.

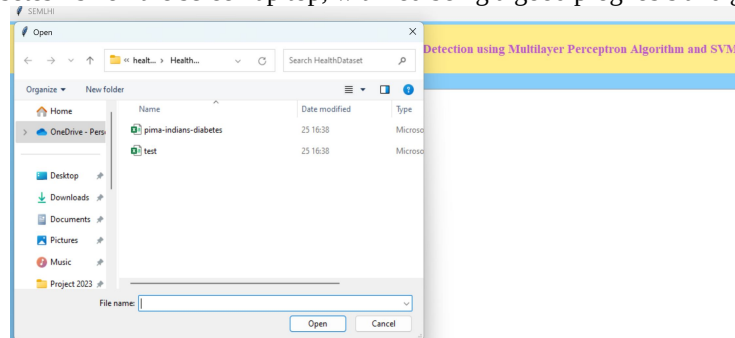


Figure 8. Load Test Set cases

The preceding screen will appear at the time of the test execution. Reporting test findings to the appropriate departments is crucial...

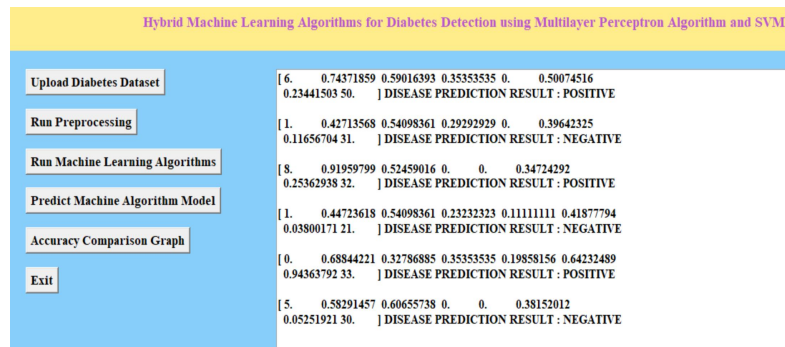


Figure 9. Output Cases

The positive and negative diabetes prediction results for the input data have been shown on the screen above.

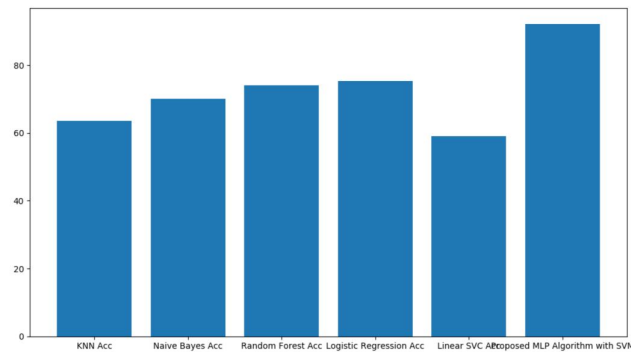


Figure 10. Accuracy Output

The MLP algorithm is one of several compared in the accompanying diagram. Using the hospital dataset, it has been proven that the extreme learning algorithm is the most effective in predicting diabetes. To better predict who may get diabetes in the future, we use an MLP and hybrid learning technique to a health dataset. As the MLP incorporates all of these features, it achieves this effect. The approach eliminates the requirement to do local minimization and run several iterations to get the minimum value. The high speed of its learning, improved generality, durability, and controllability, among other desirable features, have led to its widespread use. Many methods were explored throughout the early stages of traditional MLP development, and each one led to the creation of MLP algorithms with much improved performance. In addition, semi-supervised learning, unsupervised learning, and supervised learning have been added to the list of possible uses for the conventional MLP. These newer uses are in addition to the supervised learning applications that were already in existence.

VII. DISCUSSION

In this our main aim was to create disease prediction and detection tools based on multi-level classification (MLC) and supervised learning from medical records (MLR), to assess a freshly created MLP and SVM algorithm's performance and compare its outcomes to those of rival techniques. And multi-layer perceptron (MLP) and hybrid learning technique to enhance the algorithm based on the assessment of accuracy and the comparison with prior results. As using hybrid mode we get our Highest accuracy as in accuracy output figure we can clearly see it.

VIII. CONCLUSION

In the current investigation, we use an advanced method of machine learning to a dataset pertaining to health in an effort to improve our ability to foretell the start of diabetes. Many applications, such as batch learning, sequential learning, and incremental learning, make extensive use of the multilayer perceptron (MLP). This is due to the fact that the MLP is simple to build and furthermore benefits from rapid convergence in addition to great generalisation capabilities. This fact is the result of the interplay of a number of distinct causes. The Multi-layer Perceptron, or MLP, is a productive learning approach that was designed for use with single hidden layer feed forward neural networks. In terms of efficacy and efficiency, the MLP is an effective and efficient learning method. The area in which MLP excels the most is image recognition. In comparison to more traditional neural network methodologies,

there is a lower likelihood that problems caused by overfitting would occur, and it takes far less time to reach an expert level of competence. In addition, a less amount of information is required for training as compared to other approaches.

ACKNOWLEDGEMENT

I would like to thank Dr. Kakoli Banerjee for her support and help throughout this journey.

REFERENCES

- [1] M. Moreb, T. A. Mohammed and O. Bayat, "A Novel Software Engineering Approach Toward Using Machine Learning for Improving the Efficiency of Health Systems," in *IEEE Access*, vol. 8, pp. 23169-23178, 2020, doi: 10.1109/ACCESS.2020.2970178.
- [2] Beam, Andrew & Kohane, Isaac. (2018). Big Data and Machine Learning in Health Care. *JAMA*. 319. 10.1001/jama.2017.18391.
- [3] Ferdous, Munira & Debnath, Jui & Chakraborty, Narayan. (2020). Machine Learning Algorithms in Healthcare: A Literature Survey. 1-6. 10.1109/ICCCNT49239.2020.9225642.
- [4] Rahman, Atta & Sultan, Kiran & Naseer, Ifthikhar & Majeed, Rizwan & Musleh, Dhiaa & Gollapalli, Mohammed & Chaabani, Sghaier & Ibrahim, Nehad & Siddiqui, Shahan & Khan, Muhammad. (2021). Supervised Machine Learning-based Prediction of COVID-19. *Computers, Materials and Continua*. 69. 21-34. 10.32604/cmc.2021.013453.
- [5] Patel, Yash. (2021). Machine Learning in HealthCare.
- [6] Shams, Mahmoud & Elzeki, Omar & Abdelfatah, Mohamed & Abou El-Magd, Lobna & Darwish, Ashraf & Hassanien, Aboul Ella. (2021). Impact of COVID-19 Pandemic on Diet Prediction and Patient Health Based on Support Vector Machine. 10.1007/978-3-030-69717-4_7.
- [7] Harikumar & Mustafa, Malik & Sanchez, D. & Sajja, Guna & Gour, Sanjeev & Naved, Mohd & Jawarneh, Malik. (2021). IMPACT OF MACHINE learning ON Management, healthcare AND AGRICULTURE. *Materials Today: Proceedings*. 10.1016/j.matpr.2021.07.042.
- [8] Mienye, Domor & Sun, Yanxia & Wang, Zenghui. (2021). Improved Machine Learning Algorithms with Application to Medical Diagnosis.
- [9] Fetaji, Bekim & Fetaji, M. & Ebibi, Mirlinda & Ali, Maaruf. (2021). Predicting Diabetes Using Diabetes Datasets and Machine Learning Algorithms: Comparison and Analysis. 10.1007/978-3-030-90016-8_13.
- [10] Khan, Bilal & Naseem, Rashid & Shah, Muhammad Arif & Wakil, Karzan & Khan, Atif & Uddin, M. Irfan & Mahmoud, Marwan. (2021). Software Defect Prediction for Healthcare Big Data: An Empirical Evaluation of Machine Learning Techniques. *Journal of Healthcare Engineering*. 2021. 1-16. 10.1155/2021/8899263.
- [11] Lalmuanawma, Samuel & Hussain, Jamal. (2020). Applications of Machine Learning and Artificial Intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos Solitons & Fractals*. 10.1016/j.chaos.2020.110059.
- [12] Khan, Bilal & Naseem, Rashid & Shah, Muhammad Arif & Wakil, Karzan & Khan, Atif & Uddin, M. Irfan & Mahmoud, Marwan. (2021). Software Defect Prediction for Healthcare Big Data: An Empirical Evaluation of Machine Learning Techniques. *Journal of Healthcare Engineering*. 2021. 1-16. 10.1155/2021/8899263.
- [13] Scott, Ian & Carter, Stacey & Coiera, Enrico. (2021). Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health & Care Informatics*. 28. e100251. 10.1136/bmjhci-2020-100251.
- [14] Mijwil, Maad & Salem, Israa & Abttan, Rana. (2021). Utilisation of Machine Learning Techniques in Testing and Training of Different Medical Datasets. *Asian Journal of Computer and Information Systems*. 9. 29-34. 10.24203/ajcis.v9i4.6765.
- [15] Zhao, Rui & Yan, Ruqiang & Chen, Zhenghua & Mao, Kezhi & Wang, Peng & Gao, Robert. (2019). Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*. 115. 10.1016/j.ymssp.2018.05.050.