Turkish Journal of Computer and Mathematics Education *Vol.12 No.13 (2021), 7866-* 7871

Research Article

A Survey on various Machine Learning Approaches for thalassemia detection and classification

Muna Qais Mohammed[1] ,Jamal Mustafa Al-Tuwaijari[2]

[1,2]Department of Computer Science, College of Sciences, University of Diyala, Iraq

Email [1] :scicompms14@sciences.uodiyala.edu.iq, Email [2]: dr.altuwaijari@sciences.uodiyala.edu.iq

Article History: Received: 10 May 2021; Revised: 12 June 2021; Accepted: 27 July 2021.

Abstract: Thalassemia is a genetic blood disease caused by a deficiency in the production of hemoglobin, the central protein found in red blood cells and responsible for carrying oxygen from the lungs to the parts of the body. Hemoglobin consists of two types of protein chains, two alpha chains and two beta chains. In diagnosing thalassemia, doctors rely on two types of tests: a complete blood count (CBC) and a special hemoglobin test (Hemoglobin Electrophoresis). A complete blood count, or CBC, measures the amount of hemoglobin and various types of blood cells, such as red blood cells, in a blood sample. In this study present a survey of different method based on artificial intelligence to classify and detection thalassemia using the variable (parameter) of the CBC test which include RBC, HGB, MCV, HTC, HB . To distinguish between thalassemia minor alpha and thalassemia major beta patients. Decision tree, Naïve Bayes, support vector machine (SVM), and neural network classification method are used.

Keywords: Thalassemia, machine learning, Classification, Artificial Intelligence Techniques.

1. Introduction

Thalassemia is a blood disorder that can be passed on over the generations .which is also identified as Mediterranean "anemia" [1]. It is prevalent in the Mediterranean basin and is mostly inherited from parents characterized by irregular development of hemoglobin or by mutation of the parents' genes[2] . "Thalassemia" is derived from two Greek words: "Thalassa" which means "sea" and "Haema" which means "blood". Because of its high prevalence in Mediterranean nations, it was so named [3].Previously, the spread of thalassemia in the so-called 'thalassemia belt' was primarily limited to areas in the Mediterranean (Italy, Greece, Turkey, and Cyprus), and through the Middle East from Southern Asia to Southeast Asia (India, Vietnam, and Cambodia) [4]. Thalassemia is one of Indonesia's most common chronic disease [5]. Thalassemia is a condition with a genetic disorder that lacks its normal hemoglobin structure and the body produces impractical hemoglobin consisting of two forms, Alfa and Beta [6]. In beta-thalassemia, the globin chain is weakened or defective hemoglobin is involved, while in alpha-thalassemia, the Alfa-globin gene affects the globin chain[7]. Medically, thalassemia is divided into three types: thalassemia major, thalassemia intermediate, and thalassemia minor. The patient has severe anemia and will need blood transfusions for the remainder of their lives if they have thalassemia major [8]. Thalassemia intermedia is a form of anemia that causes mild to moderate anemia and sometimes requires blood transfusions. Patients with thalassemia minor, on the other hand, rarely need blood transfusions and tend to be in good health[9]. Thalassemia was rated that about 1.5 percent of the universal (80 to 90 million people) are beta-thalassemia carriers, with about 60,000 individuals born annually, in developing countries the largest proportion is. In 2013, thalassemia contributed to 25,000 deaths, down from 36,000 deaths in 1990[10] . The primary clinical test for the diagnosis of thalassemia is a complete blood count (CBC) [11]. People who seem to be well should have a medical checkup to see if they have thalassemia. Artificial intelligence includes machine learning as a subset (AI) that involves algorithmic methods for allowing machines to solve problems without the need for complex computer programming[12]. To mine knowledge from medical records, machine learning methods have been commonly used. Classification (e.g., is this particular patient sick or healthy) is a supervised method of learning in Machine Learning that can be used to create models that describe essential data groups. These machine learning approaches can assist researchers and doctors in making medical decisions, as well as provide answers to relevant and related health-care questions[13]. supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are the four types of machine learning models[14]. The learning algorithms that are most commonly used are Support Vector Machines (SVM) ,logistic regression(LR) ,naive Bayes(NB)

,linear discriminant analysis(LDA), decision trees(DT), k nearest neighbor algorithm (KNN) ,Neural Networks (Multilayer perceptron) ,Convolutional Neural Network(CNN). These algorithms are both a Machine Learning and an Artificial intelligence branch [15]. With the development of artificial intelligence, various types of machine learning algorithms have been developed, which will aid in improving the quality and accuracy of thalassemia disease detection and classification. As a result, early detection of thalassemia disease is important because it will assist in early treatment and recovery of the disease. It's also difficult to diagnose with high precision in the early stages of the disease [16].

2. MACHINE LEARNING

ML is the science that helps computers to learn and predict from their experiences without needing to be specifically programmed. A computer program is said to have learned if it can boost its efficiency by learning from previous experience. Machine learning, rather than AI, is more focused on data processing. Machine learning employs algorithms that allow computers to learn from data in an iterative manner. ML has advanced to a new stage in recent decades. Self-driving vehicles, successful web search, human voice recognition, image recognition, and many other applications have all been made possible by machine learning[17].Machine learning types include supervised, unsupervised, semi-supervised, and reinforcement learning[18]. Humans must practice by providing inputs and desired outputs in supervised learning [19]. Unsupervised learning is the polar opposite of supervised learning, in which the learner is left to their own devices without any branded responses. In supervised and unsupervised learning, data is either labeled or unlabeled, whereas semi-supervised learning uses both labeled and unlabeled data for preparation. Reinforcement learning algorithms learn which behaviors yield the highest reward through a trial and error process [20]. This algorithm is both an ML and an AI branch. The developed model is used in the prediction process of ML, and new data is fed into it. Using the ML algorithm, the predicted data will be available.

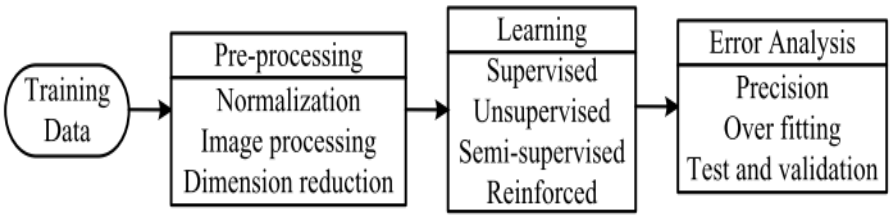

Fig 1 Machine learning (ML) process's learning phase

3. LITERATURE REVIEW

Many authors have worked on disease categorization systems in medical centers utilizing various data mining approaches and machine learning algorithms, with the goal of establishing an accurate automatic diagnosis of diseases and for better decision-making in medical institutions.The Das et al [21]. In this study a novel decision support system was proposed. The Department of Hematology of the Postgraduate Institute of Medical Education and Research (PGIMER) in Chandigarh , India, collected clinical data. The results showed 91.74 % for BTT. The Laengsri et al[22]. in this paper ,In retrospect, the authors gathered laboratory HMA data found in Thai adults 186 patients. To create a discriminant model, five machine learning methods, including (k-NN), (DT), (RF), (ANN) and (SVM). halPred achieved prediction results with an external accuracy, MCC and AUC of 95.59, 0.87 and 0.98, respectively. The Rustam et al[23]. The researchers will compare the performance of the thalassemia data classification with Fuzzy C-Means, Fuzzy Kernel C-Means, and Fuzzy Kernel Robust C-Means in this report. The researcher uses Indonesian thalassemia data with100 sample .the results show that FCM 's accuracy is 100% when training data is 90%, FRCM is 100% when training data is 90%, and FKRCM, which is the updated Fuzzy, is 100% when we use the Sigma = 0.0001 and 80% & 90% data on instruction.The Ayyıldız& Arslan Tuncer [24] proposed a paper based on different classification technique including (SVM) and (KNN).152 patients that were diagnosed with β-thalassemia at Elazıg Public Health Center. the Neighborhood Component Analysis Feature Selection (NCA) algorithm, was used for selecting features from the datasets, and the parameters selected through NCA provided high performance 97 % Area Under the ROC curve.The AlAgha et al[25]. proposed a paper based on different classification technique namely k-neighbors (k-NN), (NB), (DT) and (MLP) neural network and The location where the data was collected was the Gaza Strip. The Navies Bayes classifier was found to have the best value in distinguishing between normal and -thalassemia carriers. This mixture has a 99.47 percent precision

and a 98.81 percent sensitivity.The Roth et al[26]. proposed a paper based on SVM technique and All of the samples of more than 22,000 samples collected with 3161 (13.6 percent) of β-thalassemia carriers were examined at the Emek Medical Centre's laboratory. The SVM's formulas had 98% sensitivity and 99.77% negative predictive value, respectively.The Yousefian et al[27]. In this research, artificial intelligence methods including two algorithms ANN and MLP are used to predict the risk of diabetes in patients with β-Thalassemia Major. the accuracy of the ANN in diagnosing patients with thalassemia and revealing diabetes will be 80.78 %. Using this intelligent device and doing pre-processing, the system's precision rate is 89.48 %.The Barnhart-Magen et al[28]. this paper proposed ANN technique to diagnosis thalassemia disease. The study included a database of 526 patients, including 185 confirmed cases of alpha and beta TM, and a control group consisting of iron deficiency anemia (IDA). The result showed that ANN achieved a sensitivity of 1 and specificity of 0.958. The Paokanta et al[29]. in this paper authors study different machine learning technique (KNN), (BNs), and (LR). The dataset were gathered from hospitals in Thailand's northwestern provinces.The accuracy percentages for KNN, BNs, and LR are 85.83, 85.04, 85.04, and 82.68, respectively. The ChidozieEgejuru et al[30]. This study predicted the risk of thalassemia in all age groups by using Supervised machine learning algorithms(NB, MLP). the predictive model for the risk of Thalassemia using the naïve Bayes' classifier displayed an accuracy of 94.12 % and the predictive model for the risk of infertility using the multi- layer perceptron displayed an accuracy of 100 . The D .Tyas et al [31]. This work proposed using MLP for thalassemia classification. The dataset consists of 725 abnormal images and 99 images of normal erythrocytes. In this study, the maximum accuracy was 98.11 %. The A. jahan et al[32]. C4.5 and NB , back-propagation type ANN proposed in this paper .the dataset from VARIANT II equipment (Sysmex Corporation, Japan) (Bio-Rad Laboratories , USA). The accuracy of the C4.5 and NB was 88.56 %, while the ANN was 85.95 %. The B. Gowtham et al[33]. This researcher in tis paper using SVM , KNN ,LR , DT and RF algorithm. The dataset from University, Hyderabad, has awarded a collaborative research project. a total of 1387 sample.the Accuracy of proposed algorithm 0.76% ,0.77% ,0.78%, 0.95% , 0.96% . The M. Hajipour et al[34]. In this work RF algorithm used. With dataset (165 sample)from Mashhad University of Medical Sciences Research Council. The Accuracy of algorithm 93.72% for test data and 94.41 for training data. The A. Ghanyismaeel et al[35]. in this paper optimal neural algorithm used for thalassemia diagnosis. The dataset related to β-thalassemia extracted about (384) records from common database in ITHALNET.the Accuracy of algorithm is 0.999601%. The T. Banirostam et al[36]. k-Nearest Neighbor and Radial Basis Network proposed in this paper with Thalassemia Zafar dataset. The Accuracy rate of RBF network on 103 training sample and 153 test samples is 81.7% and the KNN 69.12%. The Cappellini et al[37] .this paper using ANN algorithm for thalassemia detection. With dataset (180 sample) from Department of Cardiology and Cardiovascular Surgery, Niguarda Ca' Granda Hospital, Milan. The accuracy rate of the system is 92.51%. The S. Thakur et al [38]. The technique that used in this work is FUZZY INFERENCE SYSTEM (FIS) with dataset from Thalassemia Welfare Society, Bhilai (Chhattisgarh, India). This results with an accuracy of about 80 %. The M. Amin et al[39]. In this study DT, NB and MLP algorithm used for thalassemia classification. The dataset is 240 sample from A medical technologist who has received a license from Bangladesh's State Medical Faculty. The accuracy of DT is 97.16% , NB accuracy is 70.28 %, MLP accuracy is 86.5566 %. The H. Chen et al[40].the technique used in this paper is logistic regression , artificial neural network with dataset 152 sample from Beijing Municipal Education Commission, China's Science and Technology Project.The accuracy of ANN is 90% and LR is 86%.

4. CONCLUSION

This paper summarizes previous studies on the identification and diagnosis of thalassemia disease using various machine learning algorithms. This survey and analysis clearly found and observed that certain machine learning algorithms, such as Decision tree, J48, SVM, KNN, and ANN, provide better accuracy in detecting and predicting thalassemia disease. Moreover, different algorithms behave differently based on different factors. To get better prediction results, consider the situation, but most importantly, the dataset and feature selection. In addition, the paper includes a survey of various types of machine learning techniques used by various authors, with each machine learning technique having some good and bad outcomes depending on the datasets and features selected, among other factors. Through this survey, we discovered that using various combinations or hybrid machine learning algorithms can increase accuracy and efficiency, and that in the future, we can work on more parameters to improve performance over the current technique.

5. REFERENCES

[1] M. D. Cappellini and A. T. Taher, "The use of luspatercept for thalassemia in adults," Blood Adv., vol. 5, no. 1, pp. 326–333, 2021, doi: 10.1182/bloodadvances.2020002725.

[2] R. Bou-Fakhredin et al., "CYP450 mediates reactive oxygen species production in a mouse model of β-thalassemia through an increase in 20-hete activity," Int. J. Mol. Sci., vol. 22, no. 3, pp. 1–14, 2021, doi: 10.3390/ijms22031106.

[3] M. A. Jalali Far, A. Oodi, N. Amirizadeh, M. Mohammadipour, and B. KeikhaeiDehdezi, "The Rh blood group system and its role in alloimmunization rate among sickle cell disease and sickle thalassemia patients in Iran," Mol. Genet. Genomic Med., no. January, pp. 1–9, 2021, doi: 10.1002/mgg3.1614.

[4] H. Frangoul et al., "CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia," N. Engl. J. Med., vol. 384, no. 3, pp. 252–260, 2021, doi: 10.1056/nejmoa2031054.

[5] E. R. Susanto, A. Syarif, K. Muludi, R. R. W. Perdani, and A. Wantoro, "Implementation of Fuzzy-based Model for Prediction of Thalassemia Diseases," J. Phys. Conf. Ser., vol. 1751, p. 012034, 2021, doi: 10.1088/1742-6596/1751/1/012034.

[6] A. Lal et al., "Transfusion practices and complications in thalassemia," Transfusion, vol. 58, no. 12, pp. 2826–2835, 2018, doi: 10.1111/trf.14875.

[7] V. Viprakasit and S. Ekwattanakit, "Clinical Classification, Screening and Diagnosis for Thalassemia," Hematol. Oncol. Clin. North Am., vol. 32, no. 2, pp. 193–211, 2018, doi: 10.1016/j.hoc.2017.11.006.

[8] A. A. Thompson et al., "Gene Therapy in Patients with Transfusion-Dependent β-Thalassemia," N. Engl. J. Med., vol. 378, no. 16, pp. 1479–1493, 2018, doi: 10.1056/nejmoa1705342.

[9] C. D.Asadov, "Immunologic Abnormalities in β-Thalassemia," Prime Arch. Immunol., no. July, 2020, doi: 10.37247/pai.1.2020.1a.

[10] B. E. Shmukler et al., "Genetic disruption of KCC cotransporters in a mouse model of thalassemia intermedia," Blood Cells, Mol. Dis., vol. 81, no. September 2019, p. 102389, 2020, doi: 10.1016/j.bcmd.2019.102389.

[11] A. N. Saliba et al., "Thalassemia in the emergency department: special considerations for a rare disease," Ann. Hematol., vol. 99, no. 9, pp. 1967–1977, 2020, doi: 10.1007/s00277-020-04164-6.

[12] K. Y. Ngiam and I. W. Khor, "Big data and machine learning algorithms for health-care delivery," Lancet Oncol., vol. 20, no. 5, pp. e262–e273, 2019, doi: 10.1016/S1470-2045(19)30149-4.

[13] R. Buettner and M. Schunter, "Efficient machine learning based detection of heart disease," 2019 IEEE Int. Conf. E-Health Networking, Appl. Serv. Heal. 2019, no. October, 2019, doi: 10.1109/HealthCom46333.2019.9009429.

[14] A. Shatte, D. Hutchinson, and S. Teague, "Machine learning in mental health: A systematic scoping review of methods and applications," 2018, doi: 10.31219/osf.io/hjrw8.

[15] M. Armanur Rahman et al., "A survey of machine learning techniques for self-tuning hadoop performance," Int. J. Electr. Comput. Eng., vol. 8, no. 3, pp. 1854–1862, 2018, doi: 10.11591/ijece.v8i3.pp1854-1862.

[16] E. M. T. El-kenawy, "A Machine Learning Model for Hemoglobin Estimation and Anemia Classification Anemia Classification Module Hemoglobin Estimation Module Data Cleaning Data Preprocessing," vol. 17, no. 2, pp. 100–108, 2019.

[17] M. A. Jasim and J. M. Al-Tuwaijari, "Plant Leaf Diseases Detection and Classification Using Image Processing and Deep Learning Techniques," Proc. 2020 Int. Conf. Comput. Sci. Softw. Eng. CSASE 2020, pp. 259–265, 2020, doi: 10.1109/CSASE48920.2020.9142097.

[18] G. Waleed Naji and J. Mustafa, "Satellite Images Scene Classification Based Support Vector Machines and K-Nearest Neighbor," Diyala J. Pure Sci., vol. 15, no. 3, pp. 70–87, 2019, doi: 10.24237/djps.15.03.486b.

[19] H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari, “Cancer Classification Using Gaussian Naive Bayes Algorithm,” Proc. 5th Int. Eng. Conf. IEC 2019, pp. 165–170, 2019, doi: 10.1109/IEC47844.2019.8950650.

[20] J. M. Al-Tuwaijari and S. I. Mohammed, “Performance Evaluation Of Face Image Recognition Based Voila-Joins With SVM,” **المجلة العراقية لتكنولوجيا المعلومات**, p. 161, 2019, doi: 10.34279/0923-009-003-014.

[21] S. Purwar, R. K. Tripathi, R. Ranjan, and R. Saxena, “Detection of microcytic hypochromia using cbc and blood film features extracted from convolution neural network by different classifiers,” Multimed. Tools Appl., vol. 79, no. 7–8, pp. 4573–4595, 2020, doi: 10.1007/s11042-019-07927-0.

[22] R. Das et al., “A decision support scheme for beta thalassemia and HbE carrier screening,” J. Adv. Res., vol. 24, pp. 183–190, 2020, doi: 10.1016/j.jare.2020.04.005.

[23] S. Hartini and Z. Rustam, “Hierarchical Clustering Algorithm Based on Density Peaks using Kernel Function for Thalassemia Classification,” J. Phys. Conf. Ser., vol. 1417, no. 1, 2019, doi: 10.1088/1742-6596/1417/1/012016.

[24] H. Ayyıldız and S. Arslan Tuncer, “Determination of the effect of red blood cell parameters in the discrimination of iron deficiency anemia and beta thalassemia via Neighborhood Component Analysis Feature Selection-Based machine learning,” Chemom. Intell. Lab. Syst., vol. 196, no. August 2019, 2020, doi: 10.1016/j.chemolab.2019.103886.

[25] A. S. AlAgha, H. Faris, B. H. Hammo, and A. M. Al-Zoubi, “Identifying β-thalassemia carriers using a data mining approach: The case of the Gaza Strip, Palestine,” Artif. Intell. Med., vol. 88, no. July 2017, pp. 70–83, 2018, doi: 10.1016/j.artmed.2018.04.009.

[26] S. L. Moulton et al., “State-of-the-art monitoring in treatment of dengue shock syndrome : a case series,” J. Med. Case Rep., pp. 1–7, 2016, doi: 10.1186/s13256-016-1019-z.

[27] F. Yousefian, T. Banirostam, and A. Azarkeivan, “Predicting the Risk of Diabetes in Iranian Patients with β-Thalassemia Major / Intermedia Based on Artificial Neural Network,” no. September, pp. 23–33, 2019.

[28] G. Barnhart-Magen, V. Gotlib, R. Marilus, and Y. Einav, “Differential diagnostics of thalassemia minor by artificial neural networks model,” J. Clin. Lab. Anal., vol. 27, no. 6, pp. 481–486, 2013, doi: 10.1002/jcla.21631.

[29] P. Paokanta, “β-Thalassemia Knowledge Elicitation Using Data Engineering: PCA, Pearson’s Chi Square and Machine Learning,” Int. J. Comput. Theory Eng., vol. 4, no. 5, pp. 702–706, 2012, doi: 10.7763/ijcte.2012.v4.561.

[30] N. ChidozieEgejuru, S. Olayinka Olusanya, A. OnyenonachiAsinobi, O. Joseph Adeyemi, V. Oluwatimilehin Adebayo, and P. Adebayo Idowu, “Using Data Mining Algorithms for Thalassemia Risk Prediction,” Int. J. Biomed. Sci. Eng., vol. 7, no. 2, p. 33, 2019, doi: 10.11648/j.ijbse.20190702.12.

[31] D. A. Tyas, S. Hartati, A. Harjoko, and T. Ratnaningsih, “Morphological, Texture, and Color Feature Analysis for Erythrocyte Classification in Thalassemia Cases,” IEEE Access, vol. 8, pp. 69849–69860, 2020, doi: 10.1109/ACCESS.2020.2983155.

[32] A. Jahan, G. Singh, R. Gupta, N. Sarin, and S. Singh, “Role of Red Cell Indices in Screening for Beta Thalassemia Trait: an Assessment of the Individual Indices and Application of Machine Learning Algorithm,” Indian J. Hematol. Blood Transfus., pp. 3–7, 2020, doi: 10.1007/s12288-020-01373-x.

[33] B. P. Gowtham et al., “Prediction of Anemia Disease Using Classification Methods,” EasyChair, 2020.

[34] O. Article and G. Algorithm, “International Journal of Health Studies A Predictive Model for Mortality of Patients with Thalassemia using Logistic Regression Model and Genetic Algorithm,” vol. 4, no. 3, pp. 21–26, 2019, doi: 10.22100/ijhs.v4i3.523.

[35] A. Ghany Ismaeel, "Diagnose Mutations Causes B-Thalassemia: Biomining Method Using an Optimal Neural Learning Algorithm," Int. J. Eng. Technol., vol. 8, no. 1, 2019.

[36] I. J. Vole, F. Yousefian, T. Banirostam, and A. Azarkeivan, "Prediction of Mellitus Diabetes in Patients with Beta- thalassemia using Radial Basis Network , and k-Nearest Neighbor based on Zafar Thalassemia Datasets," no. 4, 2020, [Online]. Available: www.sweetmaxwell.org.

[37] M. Baldini et al., "The Role of Trabecular Bone Score and Hip Geometry in Thalassemia Major: A Neural Network Analysis," Br. J. Res., vol. 04, no. 04, pp. 1–9, 2017, doi: 10.21767/2394-3718.100025.

[38] S. Thakur, S. N. Raw, R. Sharma, P. Mishra, I. Engineering, and W. Africa, "International Journal of Applied Pharmaceutical Sciences and Research," vol. 1965, pp. 88–95, 2016.

[39] M. N. Amin and A. Habib, "Comparison of Different Classification Techniques Using WEKA for Hematological Data," Am. J. Eng. Res., no. 43, pp. 2320–847, 2015, [Online]. Available: www.ajer.org.

[40] H. Chen, J. Zhang, Y. Xu, B. Chen, and K. Zhang, "Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans," Expert Syst. Appl., vol. 39, no. 13, pp. 11503–11509, 2012, doi: 10.1016/j.eswa.2012.04.001.