Milestone Transactions on
MEDICAL TECHNOMETRICS
Vol. 01, Issue. 02, December 2023



RESEARCH ARTICLE OPEN ACCESS

# Anaemia Estimation for Patients Using Lasso And Ridge Regression Algorithms

**Ambika B J[1] . Nirmala S Guptha[2] . Syeda Ayesha Siddiqha[3]**

[1]Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy for Higher Education, India
[2]Department of Computer Science and Engineering, Sri Venkateshwara College of Engineering, Bengaluru, India
[3]Department of Computer Science and Engineering, HKBK College of Engineering, Bengaluru, India

DOI: **10.5281/zenodo.10255349**
Received: 18 October 2023 / Revised: 16 November 2023 / Accepted: 01 December 2023


**Abstract –** Treatment Suggested by computer opinion is valuable in medical decision-making, saves time, is more accurate, and doesn't require hiring new workers. Numerous nutritional assessments reveal that roughly 25% of people worldwide are anaemic. A machine learning regression that can accurately detect anaemic is therefore urgently needed. To recognise anaemic, it is important to know which classifier, or combination of classifiers, produces the best level of delicacy in the classification of red blood cells. To determine and compute the anaemic, we employed the Lasso and Ridge regressions. However, the Ridge classifier outperforms the Lasso regression and reaches a higher level of delicacy. Consequently, a better and more significant method should be applied to obtain the greatest degree of finesse in medicine.

**Index Terms** – Anaemic, machine learning, Lasso, Ridge, detection.

## I. INTRODUCTION

Computers have virtually Every aspect of life, including medicine, has become mechanised, and complaint opinion is no exception. Technologies that use computers to analyse each patient's sickness enable finding complaints quicker, less expensive, and easier. They facilitate decision-making by saving time and effort. Additionally, the automated pinion system is less vulnerable to variation in croaker selections. Computers may also be useful in the prevention and treatment of anaemic. According to estimates, anaemic affects almost every child under the age of five. a serious ailment that

MILESTONE RESEARCH.IN OPEN ACCESS
Electronic ISSN: **2584-072X**

disproportionately affects mothers and children. The sophistication of clinical decision-making tools enables the use of a better model. However, one should verify if or not the mower provides or performs better delicacy in decision-making timber.

Reduced levels of haemoglobin or red blood cells (RBCs) in the blood are referred to as anaemic that have negative effects on both economic and social growth. Anaemic can be caused by an iron shortage, long-term illnesses including HIV, malaria, and TB, a lack of vitamins B12 and A, cancer, and acquired conditions that interfere with the manufacture of red blood cells and haemoglobin, even though blood haemoglobin concentration is the most reliable indicator of anaemic. When it develops during gestation, anaemic results in fatigue and low productivity and may be linked to a higher pregnancy-related and neonatal mortality risk. According to the World Health Organization (WHO), 3.0 million fatalities in developing countries were a result of maternal and new born mortality. Anaemic concern Vaticination is very useful in identifying other related disorders. Anaemic complaints are categorised based on their morphology or their underlying aetiology. Based on morphology, anaemic is divided into three categories: normocytic, microcytic, and leukemic. Anaemic can fall into one of three types depending on the underlying cause: excessive blood cell destruction, excessive extravagant blood cell damage, or blood loss from normal blood.

Normocytic, microcytic, and leukemic anaemic are the three types that are separated based on morphology. Anaemic is broken down into three categories based on the underlying cause: extreme, excessively extravagant blood cell death, and blood loss from normal blood. The most reliable blood test to assess general health and recognise a number of illnesses, including anaemic, infection, and leukaemia, is the CBC test. Haemoglobin(Hb), red blood cells (RBC), haematocrit (HCT), mean corpuscular haemoglobin (MCH), mean corpuscular volume (MCV), and other factors are among the first 15 parameters in a full blood count test. How much iron is stored in the body is determined by a ferritin test. High ferritin levels can be a sign of hemochromatosis or another iron storage problem. Low ferritin levels, a sign of iron shortage, indicate anaemic. Inherited diseases are found using a molecular test called PCR.

## II. LITERATURE SURVEY

Estimating the beneficial link requires a through analysis of the scientific literature on the connection between iron shortage and physical work, including investigations on both humans and animals. Symptoms of an iron insufficiency spectrum, including serious iron-inadequacy weakness (SIDA), direct press inadequacy paleness (MIDA), and press deficiency without sickliness, was examined (IDNA) Aerobic capacity, adherence, energy effectiveness, voluntary exertion, and job output were used to evaluate work capacity. The 29 exploratory reports that were looked at a significant negative impact of SIDA and MIDA on both animal and human aerobic capacity. Towel iron deficiency may also contribute to this impact by causing a decrease in cellular oxidative capability, which is the assumed medium for this action. However, people have not yet experienced the potent mediating consequences of decreased cellular oxidative capability that have been shown in animals. However, people have not yet experienced the potent mediating consequences of decreased cellular oxidative capability that have been shown in animals. Abidance capacity was similarly decreased in SIDA and MIDA. In both the laboratory and the field, iron deficiency in people had an impact on energetic effectiveness. Anaemic and diminished oxygen delivery are most likely to blame for the lower work productivity observed in field studies. It has not yet been

established that IDNA or iron deficiency anaemic (IDA) has any negative social or economic implications. The typical elements of the impact of IDA on the work limit are important sources of strength for properly legitimising mediations to raise the status of iron to upgrade mortal capital. This may also be true for people who pass the IDNA test, whose impacts on work capacity may be less visible but whose numbers may be far higher than those who pass the IDA test.[1].

Encyclopaedic estimates of women who are fertile might have anaemic between the ages of 40 and 50. Studies in the past have shown conflicting data to support the link between maternal anaemic and intrauterine growth limitation (IUGR). We did a thorough analysis of the literature to find studies linking maternal anaemic to problems with small for gestational age (SGA) (as a deputy for IUGR). To combine relationships that were distributed according to the haemoglobin cut-offs the authors provided, a meta-analysis was carried out. We connected 12 research that reported links link SGA and maternal anaemic. The haemoglobin cut-off for the meta-analysis was less than 110 g/L for seven associations, less than 100 g/L for seven, and less than 90 or 80 g/L for five. The new-born's chance of having SGA increased by 53 per cent in the 90- or 80-g/L group, even if the condition did not significantly connect with the 110- and 100-g/L categories [pooled OR = 1.53 (95% CI: 1.24-1.87); P 0.001]. SGA issues seem to be associated with moderate to severe maternal anaemic, but not a mild case, but because to the wide variety of research, the conclusions need to be taken with a grain of salt. Additional research should be carried out utilising datasets with better-standardised delineations [2].

Maternal iron retention must be increased during pregnancy. Since pregnancy increases plasma volume and decreases haemoglobin fixation, maternal iron status cannot be evaluated just from a haemoglobin focus. women who are pregnant multiple times or who are expecting a large child experience the greatest reduction in this risk. Contrarily. Throughout pregnancy, the mean corpuscular volume does not vary noticeably, and a newborn is likely iron deficient if the haemoglobin level is less than 95 g/L and the mean corpuscular volume is less than 84 FL. During pregnancy, the mother's iron retention needs to be boosted. Because pregnancy increases plasma volume and decreases haemoglobin fixation, maternal iron status cannot be evaluated just from a haemoglobin focus. Women who are expecting multiple children or large children see the highest reduction in this risk. On the other hand, the mean corpuscular volume remains relatively constant during pregnancy, and a new born is probably iron deficient if both the mean corpuscular volume and the haemoglobin level are below 95 g/L.

Failure of the plasma volume to increase is also associated with severe anaemic (haemoglobin 80 g/L), the delivery of tiny children as a result of both growth restriction and early labour, and these conditions. Haemoglobin levels above 120 g/L are associated with a /=3-fold greater risk of preeclampsia and intrauterine growth restriction at the end of the second trimester. Low birth weight (2.5 kg) and early labour occur when haemoglobin centralization is between 95 and 105 g/L. (37 completed weeks) are most prevalent. When combined with a mean corpuscular volume >84 FL, this is often seen as an indication of pallor in pregnant women and should be regarded as favourable.[3]. To diagnose haematological data commentary, medical practitioners require a reliable vaccination approach. There is a lot of information available about instances and their underlying illnesses. Data mining is often the process of analysing data from many angles and combining it with other results to produce relevant data. It is sometimes referred to as data or knowledge discovery. One of the numerous logical tools for data analysis is data mining software. Drug users can classify the data and illustrate the relationships between it by analysing it from

a variety of distinct boundaries or perspectives. Weka is a tool for data mining. There are many machine-learning algorithms in it. It offers the facility to categorise our data using vibrant algorithms.

A significant data mining fashion with wide applications is the bracket. It organises colourful data into categories. Every aspect of our lives uses brackets. Using brackets, each item in a piece of data is categorised into one of the listed sets of clusters or categories. We are researching the colourful Bracket algorithms in this paper. The major goals of the thesis are to demonstrate how several bracket algorithms may be compared using The Waikato Ecosystem for Knowledge Analysis, or WEKA, and to determine which approach is best for users dealing with haematological data. New Croakers or cases can predict haematological data using the proposed model. Additionally, Comment created a smartphone app that may easily provide commentary on haematological data. the elegant [4]. The biochemistry blood parameters used in this study were used to develop a decision support system that will be incredibly useful to the participants and make everything easier for them in terms of a lack of iron anaemic. The system, which is based on the pattern recognition process, is operated using the decision tree's structure, which is connected to one of the data mining approaches. The system's basic specifications for haematology parameters use ferritin, serum iron, and serum iron-list capacity characteristics; at the process' conclusion, anaemic and anaemic (-) results have been computed. The projected system uses approximated data from 96 examples. The outputs of the decision support system have exactly matched [5].

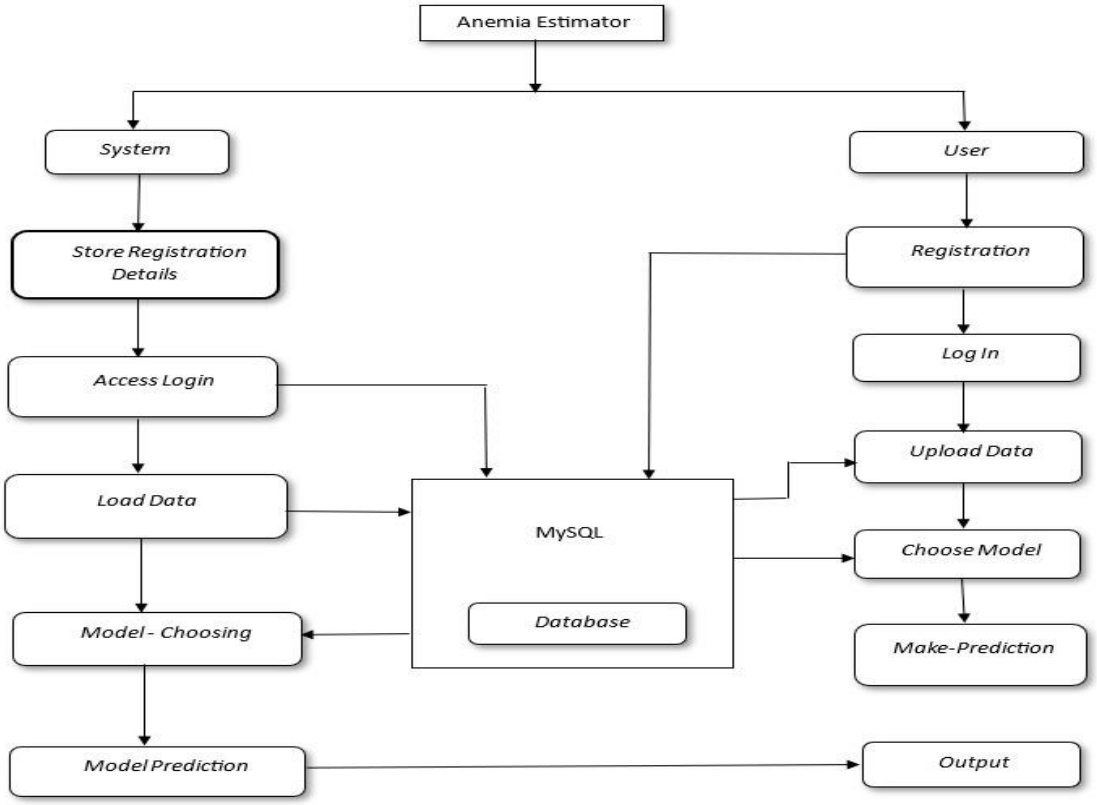


**Fig.1: Proposed system architecture**

The use of data mining in the medical field is receiving more enquiries and interest. The development of algorithms that identify and forecast knowledge from degenerating from medical environments is the focus of this novel method, known as medical data mining. The sanitarian database is one example of a medical domain where data mining is used. This database contains data that is enormous in quantity, complicated in substance, with a variety of categories, hierarchy, and of changing quality. As I mentioned before, laboratory knowledge is constantly improving and expanding. To improve conducting inquiries and evaluating reports, data mining approaches can be used to restate the unique patterns of information. The data mining bracket is dependent on data parallels. [6].

## III. PROPOSED METHODOLOFY

Due to its ability to lessen the restrictions imposed by traditional and other ways, the application is a system that has the potential to be helpful. The objective of the project is to create a rapid, precise approach to assessing and detecting anaemic. We used a robust Python environment with an algorithmic Flask framework to plan this framework.

We follow the steps below to implement the model we propose:

- Data Collection: The information needed must first be gathered for the analysis. These data may include the patient's demographics, medical background, and outcomes of laboratory tests like haemoglobin concentration, red blood cell counts, etc.
- Data Cleaning and Pre-processing: To guarantee that the data is prepared for analysis, it must be cleansed and pre-processed after it has been gathered. This includes cleaning up any incorrect or missing data, standardising the data, and dealing with outliers.
- Feature Selection: The following stage is to determine which characteristics or factors are most crucial for predicting anaemic. Numerous methods, including correlation analysis, feature ranking, and stepwise regression, can be used to do this.
- Model Training: The next step is to train the Lasso and Ridge Regression models on the data after choosing the pertinent features. To do this, separate the data into training and testing sets, choose appropriate hyper parameters for the models, and then use cross-validation techniques to optimise the models.
- Model Evaluation: After the models have been trained, they must be tested against the testing data in order to determine how well they performed. Various measures, including the coefficient of determination (R2), mean absolute error (MAE), and root mean squared error (RMSE), can be used to do this.
- Model Selection: The best model for estimating anaemic in patients must then be chosen after comparing the Lasso and Ridge Regression models.

### Algorithms used

### *A. Lasso Regression*

A statistical procedure called Lasso is used to identify points and regularise data models. Regularization is a key idea that helps prevent overfitting the data, especially when there are significant

differences between the trained and test sets of data The stylish fit formed from the training data is given a "penalty" term in order to reduce friction with the tested data and enforce regularisation. Compressing predictor variables also lessens the influence of affair variables. By utilising regression techniques, regularisation can decrease the size of sections.

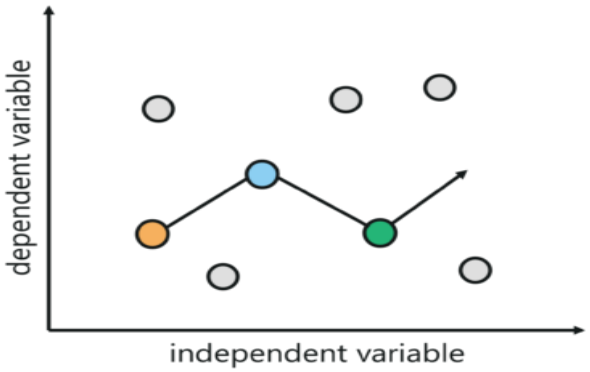


**Fig. 2. Relationship Between Dependent and Independent Variable [12]**

A regularisation method is lasso regression. It is used in place of regression techniques for a more precise vaticinator. It makes use of loss. When we have additional features, we may automatically execute point selection by using point selection. The lariat plan promotes simple models (i.e., models with smaller parameters). For models, this regression works well with considerable multicollinearity or when certain model selection processes, such variable or parameter selection, need to be automated.

$$\text{Lasso} = 1/(2*n) * \text{RSS} + \lambda|\beta| \quad (1)$$

$$\text{RSS} = \Sigma (y_i - \bar{y})^2 \quad (2)$$

where n is the number of observations, is the tuning parameter that regulates the intensity of the regularisation penalty on the absolute size of the coefficients, and RSS is the residual sum of squares. Where ȳ is the average value of the dependent variable over all observations, and yi is the dependent variable for the i-th observation (2). Lasso Regression employs L1 regularisation (will be bandied latterly in this composition). When we are more still, Lasso Regression—another name for that too used. If the L1 Regularization approach is used on a regression model. Ridge Regression would be the name if L2 regularisation was used. These will be covered in more detail in the sections that follow. With L1 regularisation, the penalty is raised by the same amount as the measure's magnitude.

Models with several sections may be affected by this regularisation type. It's possible that some segments will reach 0 and be dropped from the model. Higher penalties have an impact on measure values that are near to zero (ideal for producing simpler models). L2 regularisation, however, has no impact on any model or part elimination. lasso Regression is consequently simpler to interpret than ridge.

### *B. Ridge Regression*

A model tuning method called ridge regression can be used to analyse any multicollinearity data. In order to ensure that dissonances are significant and least-places are impartial, it applies L2 regularisation, resulting in prognosticated values that are significantly lower than factual values.

$$\text{Ridge} = 1/(2*n) * RSS + \lambda\sum(\beta\hat{}2) \quad (3)$$

where $\sum(\beta\hat{}2)$ is the sum of squares of the regression coefficients, n is the observations number, and $\lambda$ is tuning parameter that controls the strength of the regularization penalty on the squared size of the coefficients. L2 regularisation entails estimating friction that is not taken into account by the general model using the lambda function in the equation.Once the data is prepared, a method for designating it as L2 regularisation output can be used.

## IV. IMPLEMENTATION

***Technologies used.***

**1) IDE** - Popular Python programming IDE PyCharm was created to increase output and improve processes. Code completion, debugging tools, version control integration, and support for web frameworks like Django and Flask are just a few of the features it provides..

**2) Libraries**

- Pandas - It is a Python data manipulation and analysis module that offers data structures for effectively storing and handling huge datasets. It offers tools for data cleaning, merging, filtering, reshaping, and visualisation as well as the ability to handle missing data and time series data.

- NumPy- This Python package for numerical a high-performance multidimensional array object and methods for manipulating these arrays are provided by computation. It offers a wide range of mathematical operations and functions for working with arrays, including linear algebra, and is frequently used in scientific computing and data processing.

- Scikit-Learn - It is a well-liked Python machine learning package that is opensource. It offers effective tools for pre-processing data, choosing features, choosing models, and evaluating them, as well as a variety of supervised and unsupervised learning techniques. The library is made to be user-friendly, effective, and simple enough for non-techies to use while still offering sophisticated features for seasoned users.

**3) Flask Framework –** It is a simple Python web framework that offers resources for rapidly and easily creating web apps. With a basic approach to web development, it is made to be simple to use and versatile. Flask is frequently used to build RESTful APIs and microservices and allows extensions that can give your application more capabilities.

**4) MySQL Database –** Data is stored, managed, and retrieved using an open-source relational database management system (RDBMS). It has a broad range of features for building, editing, and querying databases and is based on the Structured Query Language (SQL). Databases are referred to in MySQL as groups of tables, each of which has a specific set of columns and rows that represent a certain kind of data.

***System side***

- Dataset- Several data sets will be sent to the system, and this data will be utilised to train it.
- Split - The dataset will be divided into proportions that the client will supply. Here, the data will be split into two groups: a testing group and a training group.
- Model Training - During model training, the final output will be computed using two models. Here, the user can pick a model to apply to the calculation of the outcome.
- Result - The outcome will be displayed.

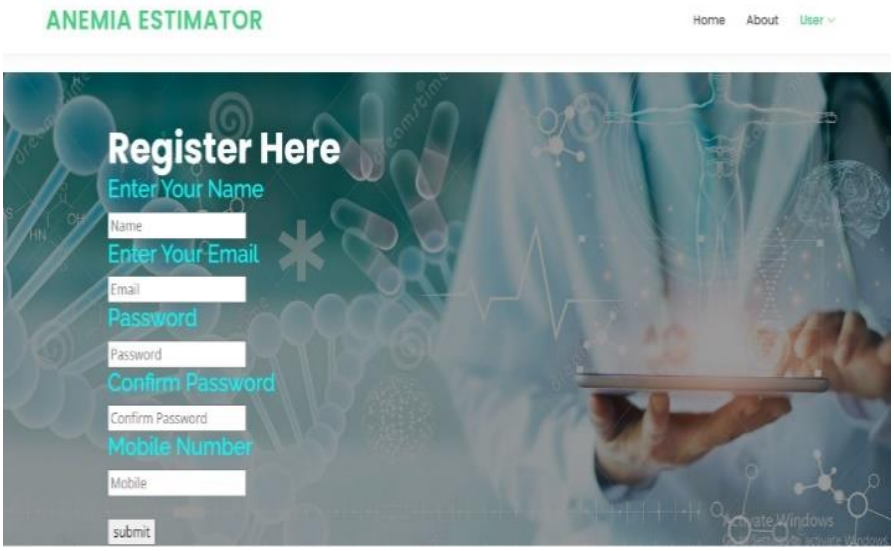


**Fig.3: Registration page setup on Local host (8080 port)**

| S.no | Age | Sex | RBC | PCV | MCV | MCH | MCHC | RDW | TLC | PLT | HGB |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **nan** | nan | nan | Red Blood Cell count | Packed Cell Volume | Mean Cell Volume | Mean Cell Hemoglobin | nan | Red Cell Distribution width | White Blood Cell (WBC) | Platelet | Hemoglobin |
| **1.0** | 28.0 | 0.0 | 5.66 | 34 | 60.1 | 17 | 28.2 | 20 | 11.1 | 128.3 | 9.6 |
| **2.0** | 41.0 | 0.0 | 4.78 | 44.5 | 93.1 | 28.9 | 31.0 | 13 | 7.02 | 419 | 13.8 |
| **3.0** | 40.0 | 1.0 | 4.65 | 41.6 | 89.5 | 28.8 | 32.2 | 13 | 8.09 | 325 | 13.4 |
| **4.0** | 76.0 | 0.0 | 4.24 | 36.7 | 89.6 | 26.7 | 30.8 | 14.9 | 13.41 | 264 | 11.3 |
| **5.0** | 20.0 | 1.0 | 4.14 | 36.9 | 89.1 | 27.8 | 31.2 | 13.2 | 4.75 | 196 | 11.5 |
| **6.0** | 24.0 | 0.0 | 4.29 | 40.1 | 93.5 | 29.6 | 31.7 | 14.5 | 13.96 | 233 | 12.7 |
| **7.0** | 28.0 | 1.0 | 4.98 | 42.3 | 84.9 | 24.9 | 29.3 | 16.2 | 9.33 | 213 | 12.4 |

**Fig.4: Patient data set containing a range of parameters.**

***User side***

- Registration - The customer needs to sign up for the organisations that were requested on the enrolment page.
- Login - Using the details they gave upon registration, such as their email address and password, the user can then access the page at any time..
- Splitting and Selecting Model - To show the dataset's findings, user can choose the desired split rate and the necessary model. In actuality, user should pick any of the offered algorithms.
- Prediction - The system can be used by the user to forecast results by entering tested values into the necessary fields. After that, the user can assess whether anaemic has harmed them.
- Logout - By logging out, the user may leave the system.

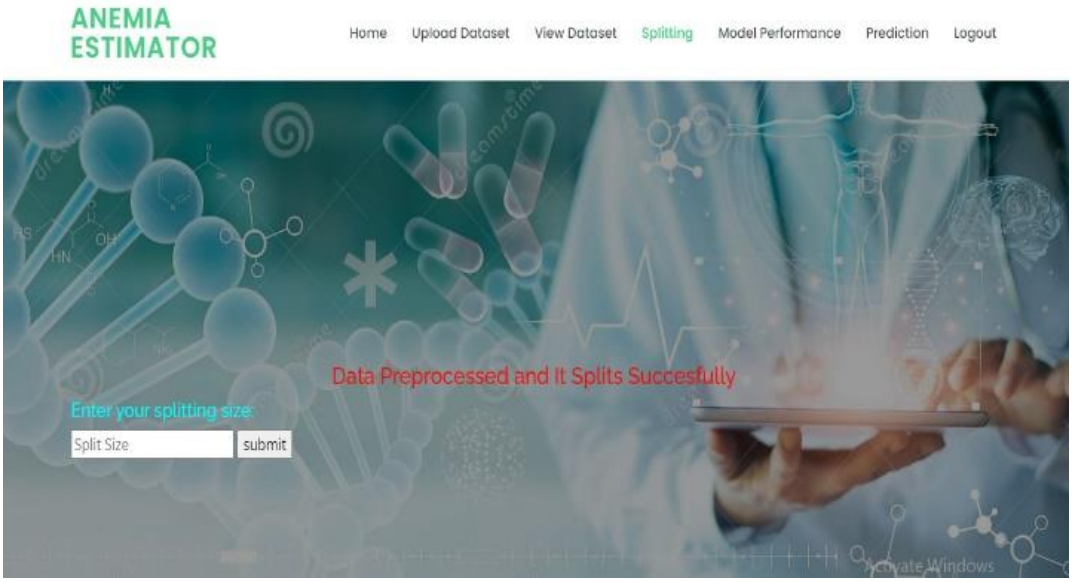


**Fig. 5: Splits the data into train and test sets before training our model.**

## V. RESULTS

This paper overcomes feature scaling, labelled data, limited interpretability, and overfitting. The problem of using the KNN algorithm and classifiers in the earlier method was considering more space to compute the data, resulting in longer training and prediction times, sensitive to the number of neighbours concerning the k value. The accuracy is 55 per cent due to problems with complex models and computation of data. Using Lasso and Ridge algorithms can help prevent overfitting and produce a more accurate model by regularizing the coefficients of less important features towards zero. The performance of the Lasso and Ridge regression models can be evaluated using metrics such as mean squared error (MSE) or R-squared. A high R-squared value or a low MSE value would indicate that the model is performing well in terms of lasso regression accuracy 84.4 per cent and ridge regression accuracy 94.5 per cent.

[Fig.6] Values of Age, RBC count, PCV count, MCV count, MCH count, MCHC count, RDW count, TLC count, and Platelets are entered in the fields to check whether a person is affected with anaemic

or not. If a person is affected by anaemic output shows a person is affected with anaemic. Or else it shows a person is not affected with anaemic.

**Table 1: Comparison table of existing algorithms and proposed method**

| Reference Paper | Algorithm | Accuracy % |
|---|---|---|
| Predicting Anaemic in Haemodialysis Patients Using Machine Learning Techniques [10] | KNN and classifiers | 55 |
| Proposed method | Lasso regression algorithm<br>Ridge regression algorithm | 84.4<br>94.5 |

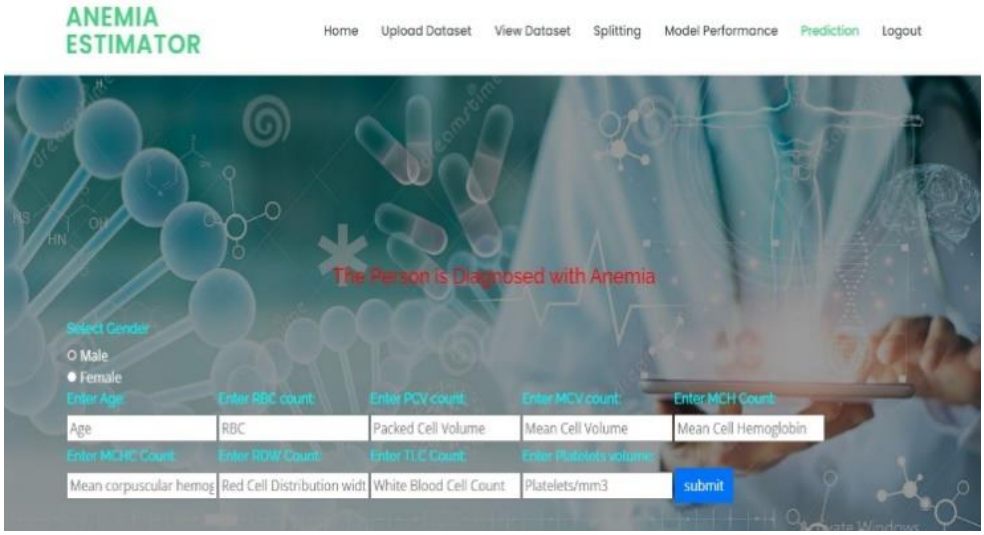


**Fig. 6: Prediction of anaemic**

## VI. CONCLUSION AND FUTURE WORK

The use of computer-based opinions in medical decision-making can be a valuable tool, saving time and producing more accurate results. In particular, the detection of anaemic is an urgent need for many people worldwide, and machine learning regression can provide an effective solution. Our study evaluated two regression methods, Lasso and Ridge, to determine the best classifier for detecting anaemic in red blood cells. Our results suggest that the Ridge classifier outperforms the Lasso regression in terms of sensitivity, making it a more significant method for medical diagnosis. Further research can build on this study to refine the classifier and improve its effectiveness in detecting anaemic, ultimately improving patient care and outcomes.

The capability of evaluating the effectiveness of adding more traits or data sources, such as genetic data, electronic health records, or data from wearable technologies. This might contribute to increasing the model's accuracy and making forecasts for specific patients more individualised. To further enhance the model's prediction performance, it may be worthwhile to investigate the use of the most

sophisticated machine learning techniques, such as deep learning. Finally, it might be beneficial to carry out additional validation experiments to judge how well the model can be applied to various patient demographics or clinical contexts.

## REFERENCES

1. Haas, J. D., & Brownlie IV, T. (2001). Iron deficiency and reduced work capacity: a critical review of the research to determine a causal relationship. *The Journal of nutrition*, *131*(2), 676S-690S.
2. Kozuki, N., Lee, A. C., & Katz, J. (2012). Child Health Epidemiology Reference G. Moderate to severe, but not mild, maternal anemia is associated with increased risk of small-for-gestational-age outcomes. *J Nutr*, *142*(2), 358-62.
3. Steer, P. J. (2000). Maternal hemoglobin concentration and birth weight. *The American journal of clinical nutrition*, *71*(5), 1285S-1287S.
4. Amin, M. N., & Habib, M. A. (2015). Comparison of different classification techniques using WEKA for hematological data. *American Journal of Engineering Research*, *4*(3), 55-61.
5. Dogan, S., & Turkoglu, I. (2008). Iron-deficiency anemia detection from hematology parameters by using decision trees. *International Journal of Science & Technology*, *3*(1), 85-92.
6. Abdullah, M., & Al-Asmari, S. (2016). Anemia types prediction based on data mining classification algorithms. In *Communication, management and information technology* (pp. 629-636). CRC Press.
7. Veluchamy, M., Perumal, K., & Ponuchamy, T. (2012). Feature extraction and classification of blood cells using artificial neural network. *American journal of applied sciences*, *9*(5), 615.
8. Bashir, S., Qamar, U., Khan, F. H., & Javed, M. Y. (2014, December). An efficient rule-based classification of Diabetes using ID3, C4. 5, & CART ensembles. In *2014 12th International Conference on Frontiers of Information Technology* (pp. 226-231). IEEE.
9. Sathiyamoorthi, V., Ilavarasi, A. K., Murugeswari, K., Ahmed, S. T., Devi, B. A., & Kalipindi, M. (2021). A deep convolutional neural network based computer aided diagnosis system for the prediction of Alzheimer's disease in MRI images. *Measurement*, *171*, 108838.
10. Swamy, R., Ahmed, S. T., Thanuja, K., Ashwini, S., Siddiqha, S., & Fathima, A. (2021, January). Diagnosing the level of Glaucoma from Fundus Image Using Empirical Wavelet Transform. In *Proceedings of the First International Conference on Advanced Scientific Innovation in Science, Engineering and Technology, ICASISET 2020, 16-17 May 2020, Chennai, India*.
11. El-kenawy, E. S. M., Eid, M. M., & Ibrahim, A. (2021). Anemia estimation for covid-19 patients using a machine learning model. *Journal of Computer Science and Information Systems*, *17*(11), 2535-1451.
12. Sreedhar Kumar, S., Ahmed, S. T., & NishaBhai, V. B. (2019). Type of supervised text classification system for unstructured text comments using probability theory technique. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(10).