REPUBLIC OF TÜRKİYE
ADANA ALPARSLAN TÜRKEŞ SCIENCE AND TECHNOLOGY UNIVERSITY

GRADUATE SCHOOL
COMPUTER ENGINEERING DEPARTMENT

VITAMIN DEFICIENCY PREDICTION USING METAHEURISTIC ALGORITHMS

YARENSU PEDÜK

M.Sc.

ADANA 2023

REPUBLIC OF TÜRKİYE
ADANA ALPARSLAN TÜRKEŞ SCIENCE AND TECHNOLOGY UNIVERSITY

GRADUATE SCHOOL
COMPUTER ENGINEERING DEPARTMENT

VITAMIN DEFICIENCY PREDICTION USING METAHEURISTIC ALGORITHMS

YARENSU PEDÜK

M.Sc.

THESIS ADVISOR
ASST. PROF. DR. ESRA SARAÇ EŞSİZ

ADANA, 2023

CERTIFICATION OF APPROVAL

VITAMIN DEFICIENCY PREDICTION USING METAHEURISTIC ALGORITHMS

This **M.Sc.** thesis, **completed under the conditions determined by the relevant regulations**, by **Yarensu PEDÜK with student number** 20801301002 in **Adana Alparslan Türkeş Science and Technology University Institute of Graduate School** Computer Engineering **Department**, has been evaluated by the following academic committee, and approved in accordance with the relevant articles of the "Adana Alparslan Türkeş Science and Technology University Graduate Education Regulations".

Prof. Dr. Ahmet BEYCİOĞLU

Institute of Graduate School Manager

Assoc. Prof. Dr. Ali İNAN

Head of the Department

Asst. Prof. Dr. Esra SARAÇ EŞSİZ

Academic Advisor

Thesis Advisory Committee Members

Asst. Prof. Dr. Esra SARAÇ EŞSİZ

Adana Alparslan Türkeş Science and Technology University

Assoc. Prof. Dr. Çiğdem ACI

Mersin University

Asst. Prof. İnayet Özge AKSU

Adana Alparslan Türkeş Science and Technology University

Thesis Defence Date

10/11/2023

DECLARATION OF CONFORMITY

In this thesis study, which was prepared following the thesis writing rules of Adana Alparslan Turkes Science and Technology University Institute of Graduate School, I declare that I provide all the information, documents, evaluations, and results in accordance with scientific ethics and moral codes without resorting to any means or assistance that would be contrary to scientific ethics and traditions. I also declare that I refer to all of the articles I used in this study with appropriate references and accept all moral and legal consequences if a situation is found contrary to my statement regarding my work.

..../..../20..

[Signature]

Yarensu PEDÜK

ÖZET

METASEZGİSEL ALGORİTMALAR KULLANARAK VİTAMİN EKSİKLİĞİ TAHMİNİ

Yarensu PEDÜK
Yüksek Lisans, Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Dr. Öğr. Üyesi Esra SARAÇ EŞSİZ
Kasım 2023, 45 sayfa

Kan, insanoğlunun varoluşundan beri büyük bir öneme sahip olmuştur. Antik çağlarda bile insanlar, hastalıkların ve sağlık durumlarının belirlenmesi için kanın önemli bir unsur olduğunu biliyorlardı. Tıp alanındaki gelişmeler ve teknolojik ilerlemeler sayesinde, kan testleri insanların sağlık durumlarını anlamak için önemli bir araç haline gelmiştir. Bu nedenle kan analizi, tıbbi teşhis ve tedavi süreçlerinin temel bir parçası haline gelmiştir. Bu çalışmada, B12 vitamini tahminlemesi üzerine bir araştırma gerçekleştirilmiştir. Veri seti, 2-92 yaş aralığında 509 kadın, 398 erkek olmak üzere toplam 907 kişinin kan değerlerini içermektedir. Tahminleme için SVR, LASSO ve Ridge regresyon yöntemleri kullanılmıştır. SVR, Lasso ve Ridge regresyon modellerinin performansını artırmak ve aşırı uyma (overfitting) problemini azaltmak amacıyla kullanılan modellere ait parametreler Grey Wolf Optimizasyon algoritması ile optimize edilmiştir. SVR için best 'C' ve gamma değerleri sırasıyla 0.072 ve 0,107 olarak bulunmuştur ve MAE, MSE, RMSE ve R^2 değerleri için optimizasyon sonrasında sonuçlar 0.145, 0.024, 0.152 ve 0.890 olarak iyileşmiştir. Lasso regresyon için best 'alpha' değeri 0.011 olarak bulunmuştur. MAE, MSE, RMSE ve R^2 değerleri için optimizasyon sonrasında sonuçlar 0.030, 0.003, 0.062 ve 0.983 olarak iyileşmiştir. Ridge regresyon için best 'alpha' değeri 0.013 olarak bulunmuştur. MAE, MSE, RMSE ve R^2 değerleri için optimizasyon sonrasında sonuçlar 0.063, 0.079, 0.289 ve 0.623 olarak iyileşmiştir. Bu çalışma ile tıp alanında kan analizlerinin önemini ve regresyon analizinin tahminleme gücünü vurgulanmıştır.

Anahtar Kelimeler: B12 vitamini tahminlemesi, Grey Wolf Optimizasyon Algoritması, Hiperparametre optimizasyonu

ABSTRACT

VITAMIN DEFICIENCY PREDICTION USING METAHEURISTIC ALGORITHMS

Yarensu PEDÜK
M.Sc., Department of Computer Engineering
Supervisor: Asst. Prof. Dr. Esra SARAÇ EŞSİZ
November 2023, 45 pages

Blood has held significant importance throughout human existence. Even in ancient times, people recognized the crucial role of blood in determining diseases and health conditions. With advancements in the field of medicine and technological progress, blood tests have become a vital tool for understanding individuals' health. Therefore, blood analysis has become an integral part of medical diagnosis and treatment processes. In this study, a research was conducted on the prediction of B12 vitamin levels. The dataset includes blood values of a total of 907 individuals, ranging from 2 to 92 years, with 509 females and 398 males. SVR, LASSO and Ridge regression methods were used for estimation. The parameters of the models used to improve the performance of SVR, Lasso and Ridge regression models and to reduce the problem of overfitting have been optimized by Grey Wolf Optimization algorithm. For SVR, the best 'C' and gamma values were found to be 0.072 and 0.107, respectively, resulting in improved outcomes with MAE, MSE, RMSE, and R^2 values of 0.145, 0.024, 0.152, and 0.890 after optimization. The optimal 'alpha' value for Lasso regression was 0.011, leading to enhanced results post-optimization with MAE, MSE, RMSE, and R^2 values of 0.030, 0.003, 0.062, and 0.983. In Ridge regression, the best 'alpha' value was determined as 0.013, and the results improved after optimization with MAE, MSE, RMSE, and R^2 values of 0.063, 0.079, 0.289, and 0.623. The importance of blood analysis in the medical field and the predictive power of regression analysis is emphosed with this study.

Keywords: B12 vitamin prediction, Grey Wolf Optimization Algorithm, Hyperparameter optimization

DEDICATION

I dedicate this thesis to my dear mother, who has never lost her efforts in starting my master's degree and preparing this thesis. I owe a debt of gratitude to Elif Kevser TOPUZ who patiently answered my questions at every stuck point, and Esra Saraç EŞSİZ, my consultant teacher, who did not give up on me in all my confusions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

MAE	: Mean Absolute Error
MSE	: Mean Squared Error
RMSE	: Root Mean Square Error
R^2	: Coefficent of Determination
SST	: Sum of Squares Total
SSR	: Sum of Squares Regression
SSE	: Sum of Squares Error
FNDDS	: Nutrient Database for Food and Dietary Research
ML	: Machine Learning
AUC	: Area Under the Curve
ROC	: Receiver Operating Characteristic
CBC	: Complete Blood Counts
GWO	: Grey Wolf Optimization
SVR	: Support Vector Regression

1. INTRODUCTION

Blood, also called the "Liquid of Life", is a red-colored liquid made up of rounds and plasmas that circulates through the veins in the human body and carry out our vital activities. In ancient times, in an ancient medical system called Humorism, it was believed that there are different fluids in the human body and that the patient must keep these fluids in balance in order to stay healthy. For this, they let the patient bleed in order to maintain the balance. It was the most common medical practice used for about 2000 years [1]. Later, it was understood that there was no suitable treatment for every disease, and its use was abandoned except for a few conditions (such as hypertension).

Research shows that medicine is an important subject in every period. B.C. Ebers Medical Papyrus [2], which is thought to have been written around 1550 and named after the Ancient Egypt expert George Maurice Ebers, who made himself known to the whole world in 1873, contains many prescriptions and magic formulas. It discusses the relationship of blood and body fluids such as urine, saliva, and feces. This source states the fact that the blood vessels completely surround the human body and the heart is the organ that governs the circulatory system, which shows how advanced medicine was at that time.

The late appearance of the Ebers Medical Papyrus and the fact that the blood still preserves its mystery caused scientists to turn to it more, and the small blood circulation was first described by Ibn-el Nafis in 1260 [3]. This paved the way for new studies. About 370 years later, William Harvey, in his 72-page work published in 1628, became the first person to describe the movements of the heart as it pumps blood [4] and then explained how the blood circulates in our body and the direction of blood flow. In order to reach this conclusion, he conducted studies on 80 animals of different species.

Diseases caused by blood loss brought blood transfusion trials, and its first documented use was in dogs in 1666 by Richard Lower. Using this work as a source, Jean Baptiste Denys conducted human blood transfusion trials in 1667 [5]. The subject in this experiment was Antoine Mauroy, who was famous as a mental patient in Paris [6]. Denys and his colleagues felt that replacing Mauroy's bad blood with good blood would be effective in curing this mental illness. They used the blood of a docile calf as good blood but were unsuccessful.

Upon Mauroy's death, blood transfusions were prohibited. The first successful human-to-human blood transfusion was performed by James Blundell in 1818 [7]. In Turkey, transfusion studies were initiated by Prof. Dr. Burhanettin Toker in 1921.

As a result of previous studies, Karl Landsteiner, who saw that animal blood and human blood was incompatible and even some human blood was incompatible with each other, as a result of his long studies, found 3 blood groups, A, B, and C, in 1901 [8]. C blood group was later changed to 0. These different blood groups found have led to an increase in people's interest in blood.

The invention of the compound microscope by Zacharias Jannsen in about 1600 made it possible to analyze blood content [9]. Jan Swammerdam was the first person to examine red blood cells under a microscope in 1658 [10]. Later, Antoni Van Leeuwenhoek described the size and shape of red blood cells, which no one had been able to study for 150 years. His work marked an important advance in the field of blood physiology by describing the presence and shape of blood cells. After the developments of the 17th century, blood analysis developed and took its current form with the development of technology.

Today, the process of analyzing the blood sample taken from our body with the help of an injector in the laboratory is called blood analysis. The blood sample is usually taken from a vein in the arm, but in some cases, a fingertip or artery may be preferred for blood collection.

The blood taken is mixed with a substance called anticoagulant. In this way, blood coagulation is prevented and the blood is properly preserved for analysis. Depending on the purpose of the test, the blood undergoes a special treatment and then the test is applied. When the test result comes out, laboratory technicians or experts evaluate the results by checking whether the results are within the normal range. The results can be used for different purposes such as disease monitoring, diagnosis, or treatment planning. Today, many diseases such as gout, kidney failure, diabetes, liver disorders, rheumatism, latent jaundice, vitamin deficiency, and infection can be diagnosed through blood tests. The fact that the results found in the tests are above or below the range values determined separately for each value as a result of long studies provides great convenience in the diagnosis of the disease. For example, a hemoglobin value with a value range of 13.5-17.5 in men and 12.5-15.5 in women indicates anemia and a higher than normal leukocyte count indicates an infection in the body.

Blood tests are an important and vital method used for the early diagnosis of congenital disorders in newborns. In the early 1960s, Phenylketonuria screening was started in newborns by looking at the level of phenylalanine in the blood by using the bacterial inhibition method by Doctor Robert Guthrie in the United States [11]. Thus, the first 'newborn screening program' began. Afterward, new ones were added to these works and they have survived to the present day. Heel prick test is used for screening in newborns. Today, more than ten diseases are screened from the heel prick test. Thanks to these scans, early diagnosis of congenital disorders that can be treated or controlled in the newborn significantly affects the health of the newborn [12]. Within the scope of the screening program, every baby has a heel prick test 48 hours after birth and is scanned by being processed into the system [13].

Another result that can be obtained from blood tests is the amount of vitamins in our body. The discovery of vitamins was revealed in a study by Casimir Funk in 1911 [14]. Since the term vitamin was not accepted by the institute he worked for, he could not include this term in his first article. One year later, he used the term vitamin in his official publication and brought it to the literature [15]. He later became the "father of vitamin therapy" and paved the way for research on vitamins.

Vitamins are organic compounds that our body needs to function properly. Since they are not produced at all or in sufficient quantities by our bodies, they are generally taken from food. But food is not the only way to get vitamins. For example, the sun is a good source of vitamin D for the human body. 13 types of vitamins are grouped among themselves and each vitamin takes on a different task in the body. If the amount of vitamins in our body is above or below the normal range, it may cause some discomfort. Since vitamins are not very stored foods, they are generally faced with diseases caused by their deficiency. To mention briefly, problems such as eye health problems in vitamin A deficiency, the disease of Beriberi in vitamin B deficiency, Pellagra, Anemia, Dermatitis and adversely affecting the development of the nervous system of the fetus during pregnancy, Gingival bleeding and tooth loss in vitamin C deficiency, rickets and osteomalacia in vitamin D deficiency, we can observe hemolytic anemia in vitamin E deficiency and extraordinary bleeding sensitivity in vitamin K deficiency.

Vitamins are divided into two groups as they dissolve in fat and water. Vitamins A, D, E, and K are fat soluble. Therefore, they are stored in adipose tissues or in the liver. Reserves of these vitamins can remain in the human body for days or months. If care is not taken,

vitamins A and D can accumulate in the body at a high rate, which can lead to the emergence of hypervitaminosis. The fat we take with food facilitate the absorption of these vitamins through digestion.

Vitamins B and C and all their subgroups are included in the class of water-soluble vitamins. Water-soluble vitamins are not stored in the human body for a long time and are excreted through urine. Since they are not stored for a long time, water-soluble vitamin sources are needed much more than fat-soluble vitamins.

Vitamin B12 is the largest and most complex vitamin in the human body [16].Vitamin B12 is a water-soluble vitamin. Therefore, it is stored in the body for a short time and then excreted with the help of urine. Vitamin B12 plays an active role in the proper functioning of the central nervous system and the production of red blood cells. Although it performs such nice functions, it must be taken in sufficient quantities. However high B12 levels can cause problems such as stress, panic attacks, nausea, vomiting, sleep disturbance, acne, abdominal pain and diarrhea. In recent years, there are studies showing that high B12 levels increase the risk of lung cancer [16, 17]. Various symptoms are observed in a huge range from neurological to psychiatric symptoms. For example, a few people with vitamin B12 deficiency may only present with classical megaloblastic anemia [18]. For this reason, most cases of vitamin B12 deficiency are misdiagnosed and overlooked in clinical practice.

Another disease caused by vitamin B12 deficiency is Alzheimer's disease. There are over 45 million patients worldwide, and their number is increasing. According to the data of the Alzheimer's Association of Turkey, if the increase continues, the number of Alzheimer's patients is expected to rise to 135 million in 2050 [19].

Alzheimer's disease, which is the most common type of Dementia, usually seen over the age of 65, is a medical condition that causes brain cells to die over time, resulting in decreased cognitive functions and memory loss. Studies have proven that vitamin B12 deficiency may be associated with Alzheimer's disease [20].

Blood tests are used not only for disease prediction but also for early diagnosis. A recent study presented a method which cancer could be diagnosed up to 4 years earlier, with a non-surgical blood test instead of traditional cancer diagnosis methods. Researchers aimed to detect early-stage cancer cases by focusing on the marks left by cancer cells in the bloodstream with this method they developed. This early diagnosis can increase the chances

of cure and prevent disease progression. This study is considered as a major advance in cancer early diagnosis and screening. With this method, it is thought that the life expectancy and treatment results of cancer patients can be improved [21].

While advances in the field of blood continued like this, the world on the other hand started to talk about the reality of Artificial Intelligence. The idea of imitating intelligent behavior and critical thinking using computers was first put forward by Alan Turing in 1950. In his book "Computers and Intelligence" he defined a test to determine whether computers have human intelligence. Today, this test is called the Turing Test. The term Artificial Intelligence was used by John McCarthy in 1956 and defined as "the science and engineering of making intelligent machines" [22].

Although Artificial Intelligence started as a simple set of rules, it soon developed into a discovery that included more complex algorithms and increasingly exhibited behavior similar to the human brain. With this rapid progress, it has made a place for itself in many areas. The medical industry is one of these areas. Studies have proven that it gives successful results in cancer diagnosis and it has achieved 93% compliance with an expert committee for breast cancer treatment [23]. Because of its late diagnosis, lung cancer is one of the leading causes of cancer-related deaths [24]. For this reason, its early diagnosis is very important. Artificial intelligence successfully predicts not only early diagnosis but also postoperative prognosis, tumor recurrence risk estimation, and two-year overall survival [25]. In another study, it achieved a 96.67% accuracy rate in lung cancer diagnosis [26].

Today, the increase in the sensitivity of robots and the gradual improvement of their mobility have enabled robots to gain a place in the medical field. It is thought that if robots work as medical personnel, the workload of the personnel will lighten, patient satisfaction will increase and the cost will decrease [27]. Due to the epidemics experienced in recent years, the idea of using robots in the medical field has gained importance [28]. There is no robot with artificial intelligence that can perform and manage the entire surgical process. Artificial intelligence and robotic technology are used to support surgeons during surgery and to optimize surgical procedures. However, the use of artificial intelligence in the field of surgery is rapidly advancing and plays a greater role in surgical processes. Artificial intelligence is used in areas such as preoperative planning, image processing, movements of robotic arms, tissue recognition and analysis of surgical results. In this way, surgeons can perform more precise, effective and safe surgeries. For example, the Da Vinci Surgical System, developed

by Intuitive Surgical, the first artificial intelligence operating robot, is a system that combines artificial intelligence and robotics [29]. While surgeons manipulate robotic arms using the control console of the Da Vinci Surgical System, artificial intelligence intelligence algorithms also offer optimizing support in surgical procedures [30]. Although it is used in many areas, it is especially preferred in areas such as prostate, gynecological, heart and kidney surgery. Today, intensive studies are carried out for joint surgery, especially in China. It is thought that robots can provide sufficient precision for robot-assisted knee replacement and hip surgeries, and even make future improvements in prosthesis placement [31]. The predictions made by artificial intelligence help us in the decision-making process [32].

Artificial intelligence, as a rapidly developing technology in recent years, has created great effects in many sectors. Especially in the field of forecasting, artificial intelligence techniques play an important role in predicting future events or values [33]. Artificial intelligence can develop models that can make predictions by extracting meaningful information from large data sets and analyzing this information.

Machine learning is a fundamental approach that enables artificial intelligence to make predictions. This approach enables a model to learn from data and predict future events using its experience. Machine learning works by analyzing data, identifying patterns, and using those patterns to predict future events [34].

Data mining is the process of finding hidden information and patterns within large data sets. Data mining involves the discovery and preparation of data sets used for forecasting methods [35]. At this stage, various data mining techniques are used to extract meaningful information from the data.

Among the prediction methods, classification, clustering, and regression models come to the fore. Classification divides the instances in the dataset into different classes and ensures that new data is assigned to these classes. Clustering groups data with similar characteristics in the dataset and reveals structures belonging to these groups. Regression predicts future values by modeling the relationship of a dependent variable to independent variables [36].

These prediction methods are used together with data mining and machine learning techniques to provide more precise and accurate predictions. For example, classification models can diagnose diseases based on specific symptoms, cluster models can optimize marketing strategies by segmenting customers, and regression models can help decision-

making processes by predicting future trends. Regression is an analysis method included in both data mining and machine learning. While data mining uses regression analysis as a tool to extract information from large data sets, machine learning models use regression analysis as a tool to make predictions.

The organization of this thesis is as follows:

In the first chapter, the subject of the thesis is introduced, and the advantages of the algorithms discussed in the thesis are explained by comparing them with similar ones. The previous works on similar prediction problems are given in the second chapter. Details of the dataset used in this thesis and methods proposed for accurate prediction are explained in the third chapter. In the fourth chapter, the proposed algorithms within the thesis are explained. The fifth chapter contains the comparison of the proposed algorithms with the traditional algorithms and the discussion of the test results. The sixth chapter covers future works and a general evaluation of the algorithms discussed.

2. LITERATURE REVIEW

Today, in the fields of medicine and data analysis, the processing of biomedical data such as blood tests with classification and regression models plays an important role. The reasons and advantages of combining blood tests with these models are of great importance for both medical and analytical approaches.

Classification models can be used to describe different medical conditions or diseases. Likewise, regression models can help predict the value of a particular biological measure based on a patient's age or health status by examining the relationship between blood components.

The advantages of combining blood tests with classification and regression models are quite diverse. First, we can harness the power of these models to understand and evaluate large volumes of complex data. They can also help us achieve more precise results in critical areas such as early detection and disease management.

Recent years have witnessed a growing body of research dedicated to leveraging machine learning techniques to provide medical support in the field of medicine. However, it is worth noting that, in some instances, finding literature sources that directly align with our work has been challenging. To address this gap, we have categorized similar studies for a more focused discussion.

The study conducted by Kirk et al. [37] used regression algorithms and clustering on the NHANES dataset to estimate the plasma vitamin C levels of 2952 American adults. The objective of this study is to predict the current data required for personalized nutritional recommendations. Variables with known or hypothesized associations with plasma vitamin C were selected for the study, while variables that were expensive or difficult to obtain were excluded. The best performance was achieved with the XGBoost regressor with random forest showing similar performance. Although clustering to divide the data into smaller homogeneous groups was explored, it proved to be unsuccessful. The low R-squared scores obtained by the models are believed to be a result of the low resolution of the NHANES data, particularly the dietary data. This underscores the need for high-quality datasets in Precision Nutrition research. As a result, the maximum cross-validated R-squared score achieved was 0.304 with the XGBoost regressor.

Studies on determining medical conditions, such as using blood tests to diagnose diseases and predicting vitamin deficiencies, further emphasize the impact of machine learning in the healthcare sector. Below, some examples of such studies will be presented. These studies play an important role in the early diagnosis of diseases, increasing compliance with treatment and making patients more effective in their health management.

Kılıçarslan et al. [38] they worked with a hybrid model in which they optimized Particle Swarm Optimization, Cat Swarm Optimization, and Particle Swarm Optimization with GWO for cardiovascular disease detection. MNIST dataset was used. 10-Fold Cross validation was applied and the closest result to reality was obtained with an accuracy value of 0.93 in the hybrid model.

Mohakud et al. [39] optimized the parameters of the CNN model with GWO to classify skin cancer. The model was tested on the International Skin Imaging Collaboration skin lesion multi-class dataset. The model gave an accuracy rate of 98.33%.

Research to classify blood tests for the coronavirus, which has become an epidemic and caused the death of many people in recent years, has become quite common. For example, Zhang et al. [40] used 38 blood test parameters from the blood tests taken on the first day of hospitalization of 422 COVID-19 patients from January 2020 to June 2021, using various machine learning models to distinguish between mild and severe cases. Different models such as the random forest model (Area Under the Curve(AUC): 0.89), pure Bayesian model (AUC: 0.90), support vector machine model (AUC: 0.86), and KNN model (AUC: 0.78), COVID-19 successfully classified his patients. The highest-performing model was the pure Bayesian model. These results provide an effective way of evaluating patients' mild or severe cases based on first-day blood tests. Similar studies are also listed below.

Formica et al. [41] have achieved 83% sensitivity and 82% specificity by developing a CBC-based machine learning model but is based on a limited sample (171 patients). Similarly, Banerjee et al. [42] studied a dataset containing 598 cases and achieved high specificity (91%) but low sensitivity (43%), suggesting that the model is not suitable for early detection purposes. In addition, Avila et al. [43] developed a Bayesian model reporting 76.7% sensitivity and specificity using the same dataset, Joshi et al. [44], on the other hand, obtained 93% sensitivity but 43% specificity with the logistic regression model in a data set of 380 cases. While these studies demonstrate the potential for the use of blood tests in diagnosing

COVID-19, they highlight the need for more comprehensive studies based on larger and consistent datasets.

Kurstjens et al. [45] aimed to develop a machine-learning algorithm to assess the risk of low body iron storage in patients with first-degree anemia, based on basic laboratory test results such as complete blood count and C-reactive protein. The total data was obtained from three different hospital laboratories, including Jeroen Bosch Hospital (n=3.797), Medlon BV Enschede (n=8.021), and St. Jansdal Hospital (n=191). Patients were selected based on complete blood count, C-reactive protein (CRP), and ferritin concentration. The goal of the study was to classify ferritin as 'low', 'not low', or 'uncertain'. Two separate machine learning algorithms were developed; one was based on data from Jeroen Bosch Hospital, and the other was based on data from Medlon BV to explain weak adaptability. The JBH-S algorithm achieved an ROC AUC of 0.92 on the validation dataset (n=1.140, Jeroen Bosch Hospital), while the Medlon-R algorithm reached an ROC AUC of 0.90 on the validation population (n=1.604).

Pan et al. [46] aimed to create a machine-learning model to examine iron deficiency anemia in women following sleeve gastrectomy in obesity surgery, beginning 12 months after the surgery. A total of 407 patients between the ages of 26-36 were followed for a period of 12 months. Feature selection was done, and features including ferritin, age, hemoglobin, creatinine, and fasting C-peptide were chosen for the SVM model. The prediction results yielded an AUC value of 0.85 for the training set and 0.79 for the validation set.

James et al. [47] conducted research on how routine blood test results and clinical data could be used to diagnose and classify heart failure using machine learning algorithms. In this study, routine blood test results obtained from 8.031 heart failure patients were used, along with an equal number of control group subjects. The data were divided into training and test sets in an 80:20 ratio. NT-proBNP was used for diagnostic comparison. Additionally, a dataset consisting of data from 698 patients, out of which 314 were diagnosed with heart failure, was utilized. This dataset was used to train and test machine learning models in a 70:30 ratio. First, the models were trained on raw routine blood test results (rawFBC), and then on a dataset that included rawFBC along with biochemistry and routine blood test results.

The results indicate that the AUROC (Area Under the Receiver Operating Characteristic Curve) values for predicting heart failure using rawFBC and biochemistry were 0.93 and 0.91,

respectively. Furthermore, the AUROC values for predictions made using routine blood test results and NT-proBNP were found to be 0.87 and 0.81 respectively. In the validation cohort, the AUROC value for predicting heart failure with rawFBC was 0.83, and the AUROC value for predicting NT-proBNP levels (≥ 34 pmol/L) was determined to be 0.97.

Guncar et al. [48] conducted a study based on laboratory blood test results to predict hematological diseases. They utilized a dataset consisting of a total of 8,233 cases and 371,341 blood test results from the University Medical Centre of Ljubljana (UMCL). From the 181 features in this dataset, 61 parameters were selected. Additionally, 20 additional cases were added to compare the diagnostic performance of doctors. For their predictions, they used Naive Bayes, Decision Tree, SVM, and Random Forest models, with the best result achieved using the Random Forest model, indicating an accuracy rate of 0.59.

Valderrama et al. [49] examined an intensive care unit (ICU) dataset collected from Canada, which included a total of 55,689 ICU admissions from 48,672 patients. The research aimed to compare the performance of two different machine learning approaches in predicting normal and abnormal results for 18 different blood laboratory tests. The first approach (Approach 1) used patients' vital signs, age, gender, and admission diagnosis as features, while the second approach (Approach 2) included the same features as well as new features based on conditional entropy and conditional probability, which was a novel approach. Both approaches evaluated the performance using four different machine learning models (fuzzy model, logistic regression, random forest, and gradient boosting trees) and ten different metrics (sensitivity, specificity, accuracy, precision, negative predictive value [NPV], F1 score,AUC, precision-recall AUC, average G, and balanced accuracy index). Approach 1 achieved an average AUC of 0.86 across the 18 laboratory tests, with a sensitivity of 78% and specificity of 84%. On the other hand, Approach 2, which included new features, achieved an average AUC of 0.89 with improved sensitivity (84%) and specificity (84%). The incorporation of these new features, especially the pretest probability feature, led to significant improvements in performance, making Approach 2 a more favorable choice for predicting laboratory test results.

Meng et al. [50] designed this study as a retrospective examination of a single-center hospital cohort. Initially, they had medical records for 5,060 coronary heart disease patients (2,365 males and 2,695 females) aged between 1 and 97 years, spanning 8 years from 2009 to 2017. Additionally, there were 5,051 health controls and 5,075 cases of other diseases. They

developed a two-layer Gradient Boosting Decision Tree (GBDT) model based on routine blood data, which was used to predict the risk of coronary heart disease. This model was able to identify 86% of individuals with coronary heart disease.

A dataset containing 15,000 routine blood test results was created, and this dataset was used to train the GBDT model to classify healthy conditions, coronary heart disease, and other diseases. According to the classification results after machine learning, the sensitivity for detecting health conditions was approximately 93% for all data, and the sensitivity for detecting coronary heart disease was 93% for data that included this disease.

There is no very similar study in the literature regarding the prediction of B12 vitamin levels using blood test results. Based on our research, the closest studies that can be closely related to this topic are presented below.

Tamune et al. [51] used machine learning models to predict vitamin B1, B9, and B12 deficiencies from blood tests. They examined 497 patients experiencing severe psychiatric episodes over a period of 2 years. KNN, Logistic Regression, and Random Forest models were trained using age, gender, and the results of 29 routine blood tests. Except for KNN, the performance of the other classifiers yielded similar results, but on average, the best AUC results for B1, B9, and B12 vitamins were 0.716, 0.796, and 0.599, respectively.

Sharifmousavi et al. [52] aimed to predict the levels of selenium, B12, and D3 vitamins using plasma samples from 99 Multiple Sclerosis (MS) patients and 81 healthy individuals. They employed three different machine learning models: Support Vector Machine, Decision Tree, and K-nearest neighbors (KNN). According to the compared results, the SVM model achieved the best accuracy (98.89%), precision (98.98%), and sensitivity (98.98%) among the models used.

3. MATERIAL VE METHODS

In this section, the data sources, research methods, and analysis techniques used in the study are clearly defined, and the reliability and reproducibility of the study are ensured.

3.1. Data Set

The data set used includes the blood values of 907 people, 509 women and 398 men, between the ages of 2 to 92. 57 unit values were calculated and tabulated for each patient. If we need to explain these values in general; demographic information such as gender, age and date is used to determine the patient's profile and is important for the correct interpretation of results. Glucose determines the blood sugar level and can be evaluated to diagnose diabetes or to control existing diabetes. HbA1c (hemoglobin A1c) is a measurement that reflects long-term glucose control and is used to monitor diabetes. HbA1c is the value of HbA1c expressed in an international unit.

Creatinine is an important marker for assessing kidney function. High creatinine levels can be a sign of kidney damage or kidney failure. Urea and uric acid are also parameters used to evaluate kidney function.

ALT (alanine aminotransferase) and AST (aspartate aminotransferase) are enzymes used to assess liver function. High levels may indicate liver damage or disease. ALP (alkaline phosphatase) is used to assess the health of the liver and biliary tract. LDH (lactate dehydrogenase) is a marker that indicates cell damage and the condition of tissues. GGT (gamma-glutamyl transferase) helps detect liver and biliary tract diseases.

Albumin and bilirubin are components used to assess liver function and detect conditions such as kidney failure. Electrolytes such as sodium, potassium, calcium, phosphorus and magnesium indicate the mineral balance in the body and are important for maintaining normal functions.

Total cholesterol, HDL (good) cholesterol, LDL (bad) cholesterol, and triglycerides are components of the lipid profile used to assess cardiovascular disease risk. Iron and its iron binding capacity are used to diagnose anemia and monitor response to treatment.

CK (creatine kinase) is an enzyme that indicates muscle damage and some muscle diseases. ASO (anti-streptolysin O) is an antibody test used to detect streptococcal infections. CRP (C-reactive protein) is a marker of inflammation.

WBC (white blood cells), RBC (red blood cells), HGB (hemoglobin), HCT (hematocrit), MCV (mean erythrocyte volume), MCH (mean erythrocyte hemoglobin), and PLT (platelets) show the amount and characteristics of blood cells. These parameters are important in the diagnosis of infections, anemia, blood diseases, and other health problems.

MPV and PDW can be helpful in detecting coagulation disorders by assessing the size and distribution of platelets. PCT is used to determine how much platelets are in the body. The counts and percentage of lymphocytes, neutrophils, monocytes, eosinophils, and basophils help evaluate immune system functions. It is used to monitor conditions such as sedimentation rate, inflammation and infection.

Also, a complete blood count can be used to detect certain hormonal disorders. T3, T4, and TSH are measured to evaluate the function of the thyroid gland. Ferritin shows the status of iron stores in the body, while vitamin B12 shows whether there is enough in the body.

Table 3.1. Vitamin B12 value ranges of women in the data set

Gender	#Null	#<214	#214<x<914	#>914	Total
Woman	30	34	422	23	509

Table 3.2. Vitamin B12 value ranges of men in the data set

Gender	#Null	#<197	#197<x<866	#>866	Total
Man	10	58	326	4	398

In Table 3.1, vitamin B12 values in the data set were categorized according to age and value ranges. Of the 907 people in the dataset, 509 of them are women and 30 of them contain null values. Of the remaining people, 34 had vitamin B12 deficiency and 23 had vitamin B12 excess. Vitamin B12 values were found in the normal range in 422 of them. In Table 3.2, there are 398 men and 10 of them contain null values. 58 of them have vitamin B12 deficiency and 4 of them have vitamin B12 excess. It is observed that 326 of them have normal vitamin B12 values.

Table 3.3. Distribution of Vitamin B12 value ranges of people in the data set by age groups

Age Ranges	#Null	#<200	#200<x<900	#>900
0-25	5	25	138	0
25-45	10	22	219	7
45-60	12	18	226	5
60-75	11	16	145	11
75-90	2	3	26	4
90+	0	0	2	0
Total	40	84	756	27

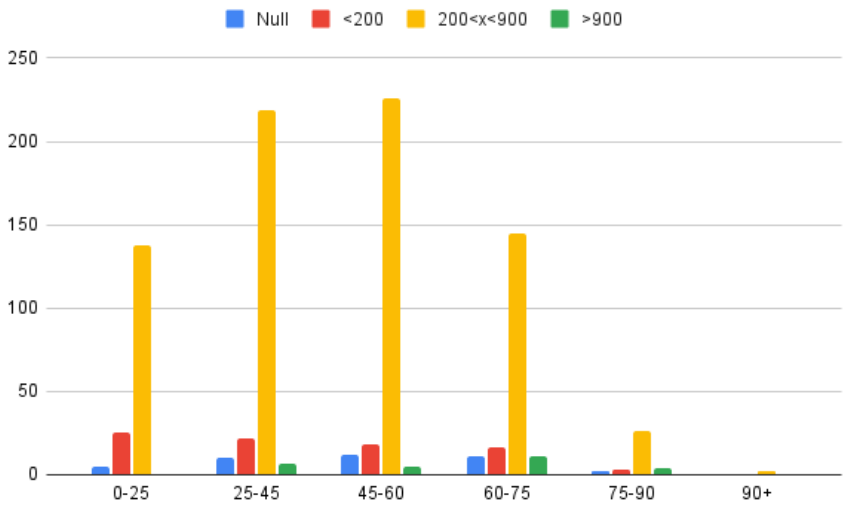

Figure 3.1. Distribution of Vitamin B12 value ranges of people in the data set by age groups

In Table 3.3.and Figure 3.1. of the same table is also included. In our data set, which included 907 people, vitamin B12 value was not found in 40 people, but vitamin B12 deficiency was found in 84 people and excess in 27 people. B12 values of 83.35% (756 people) of the whole data set are considered normal.

Since the data set used in the project is numerical, date and gender columns of various data types that caused problems during implementation were removed from the data set. There is

empty data in the dataset. Filling some cells with a prediction approach can have a negative impact on error counts because there are too many cells containing empty data. Thus, first of all, empty rows (40 rows) in column B12, which is the target value, were removed from the data set. Afterward, rows with more than 80% null values were removed from the dataset. After these operations, the data set size from 900X60 was reduced to 148x59. Some values have an int data type, while others have a float data type. Values with data types other than float are converted to float data types as they will cause data loss.

When determining the X (argument) value, all rows of the data set except the last column are selected and this data is assigned to the variable named "X" as a NumPy array. Similarly, when specifying y (dependent variable), all rows and the last column of the dataset (dependent variable) are selected, and these data are assigned to the variable "y" as a NumPy array. After this step, assignment techniques are applied and iterative assignment gives good results. But on the other hand, the feature selection didn't work very well.

3.2. Imputation

Imputation is the process of determining the origin or identity of something in general. This term refers to the attribution of an event, a work, or a piece of information to a specific source or person. Imputation is the process of determining the accuracy and reliability of a piece of information and is widely used in many fields such as researchers, historians, archaeologists, and scientists.

The history of imputation dates back to ancient times. This concept, which first appeared in Ancient Greece, played an important role in philosophy and law. Aristotle discussed the concept of imputation in his studies on ethics and political philosophy and used imputation to determine who is responsible for an action. In the ancient Roman period, imputation was used in legal systems to associate a person with a crime as guilty or innocent.

The concept of imputation in the modern sense gained importance in the field of social sciences and statistical analysis at the beginning of the 20th century. In this period, imputation methods were developed to determine the causes of a result in statistical analysis and to attribute sources. Imputation is also known as a statistical method used to replace missing or erroneous data. Imputation is a common practice especially in order to eliminate the deficiencies in the survey data and to interpret the statistical results correctly.

Imputation has many different uses. Especially historians and archaeologists use imputation methods to determine the authenticity of ancient remains or documents. In this way, historians can accurately understand past events and correct misinformation about the past. Similarly, art historians use imputation to identify the creator of a work of art or work. In short, imputation is a widely used tool to determine the accuracy of sources and information [53].

Imputation has some advantages. First, imputation is necessary to make decisions based on accurate and reliable information. Analyzes and results without considering incomplete or incorrect data may be misleading [54]. Thanks to the implantation, missing or erroneous data is replaced and more reliable results are obtained. In addition, imputation can improve the quality of scientific studies by reducing data loss[55]. Statistical analyses are based on more solid foundations, resulting in more reliable results.

However, imputation also has disadvantages. Imputation methods are based on some assumptions in data analysis and these assumptions may not always be valid in the real world. As a result of wrong assumptions, wrong conclusions can be reached. Also, the imputation process can be time-consuming and complex. In order to make an imputation correctly, it is necessary to choose the appropriate methods and apply these methods correctly. Incorrect or erroneous imputation results can affect the reliability and validity of the analyses.

In conclusion, imputation is a method used to identify the source of information and evaluate its accuracy. Historians, archaeologists, statisticians and many scientists aim to obtain more accurate results by using imputation in their studies. While the advantages of imputation enable making decisions based on accurate and reliable information, the disadvantages can be false assumptions and a time-consuming process. Although imputation is an important tool in scientific research, it is necessary to choose the right methods and apply them carefully.

There are several reasons for using imputation in my study. First and foremost, the presence of missing data in my study can jeopardize data integrity and significantly reduce the reliability of the analysis. By filling in missing data, we maintain the consistency of the dataset and help obtain more robust results. Additionally, since my study involves statistical methods like regression analysis, there is a high potential for missing data to negatively impact regression predictions. Filling in missing data improves the accuracy of regression models, enhancing the reliability of the analysis. This also reduces potential bias effects caused by missing data.

A dataset with missing data can lower the statistical power of an analysis. Imputation increases the power of the analysis by using more data points. Finally, it's important to note that missing data can increase the risk of plagiarism. By filling in missing data, we preserve data integrity and minimize the risk of plagiarism. For all these reasons, the presence of missing data in my study is a significant issue that necessitates the use of imputation. Imputation enhances the reliability of data analysis and positively impacts the success of the study.

Iterative imputation stands out as one of the most suitable options to address missing data in my study. Iterative imputation offers a versatile approach by using regression or other statistical models to predict missing data. This allows it to handle complex relationships and interactions between variables in the dataset effectively. In this study, missing data may have occurred for specific reasons. Iterative imputation offers a process that includes error checks and multiple iterations to address such special cases. This helps minimize the bias effects caused by missing data. Iterative imputation uses multiple pieces of data to predict missing data, enhancing the statistical power of the analysis. This results in a stronger and more reliable analysis.

Iterative imputation is an effective method for handling the complexity of filling in missing data in my study. For these reasons, this method have chosen to minimize the impact of missing data on the study's results and obtain more reliable outcomes. The iterative imputation process consists of the following steps:

- **Initial Prediction:** Other variables around missing data are used to predict missing data. Multiple regression analysis or another statistical model is often used. The initial predictions are used to fill in the missing data.
- **Interactions Between Variables:** Iterative imputation can also consider interactions between variables. This can help make better predictions for missing data.
- **Completed Data:** At the end of the first iteration, a dataset is created with predictions for missing data. However, these predictions may not be accurate and may still contain errors.
- **Error Control:** An important stage of iterative imputation is adding a mechanism to control and correct prediction errors. This helps improve the accuracy of predictions.

- **Iterations:** Iterative imputation can include a cycle where predictions are continually updated. Initial predictions may be incorrect, so after several iterations, better predictions are achieved.
- **Consolidation of Results:** At the end of iterative imputation, all the predictions made to fill in missing data are compiled and combined. This creates the final complete dataset.

3.3. Regression

Regression analysis is a statistical method. It is used to examine the relationship of a dependent variable with one or more independent variables, to model the relationship between variables, and to predict future values. Regression analysis is a widely used research method in many disciplines, especially in social sciences and economics. In regression analysis, the dependent variable represents the result of the event or process to be analyzed, while the independent variables represent the factors affecting this result. The purpose of regression analysis is to create the best model that can explain the variation of the dependent variable. This model is called the regression equation and is used to estimate the dependent variable.

There are two main types commonly used in regression analysis: simple regression and multiple regression. Simple regression examines the relationship of a dependent variable with only one independent variable, while multiple regression evaluates the effect of more than one independent variable. These methods are used depending on the data set and research question.

The regression analysis has many advantages. First, this method provides the possibility of expressing the relationship between variables quantitatively. In this way, information about how a dependent variable is affected by a particular independent variable and the magnitude of this effect can be obtained. Also, regression analysis can be used to predict future values of the dependent variable. These forecasts play an important role in the planning, decision-making, and strategy-making processes. Also, regression analysis has the potential to influence the results of outliers or abnormal observations in the data set. Identifying and managing these outliers is important.

In conclusion, regression analysis is a statistical method that examines the relationship of the dependent variable with one or more independent variables and makes predictions. This analysis is used in many fields, from social sciences to economics, from marketing research to

epidemiology. The advantages of regression analysis are its ability to quantify the relationship between variables and to predict future values. However, there are points to consider, such as correct interpretation and management of outliers. Regression analysis, when applied correctly, can provide researchers with valuable information and contribute to their decision-making processes.

However, regression analysis also has some disadvantages. First, regression analysis does not imply that the relationship between variables is causal. An observed relationship between two variables may be based on secondary or coincidental factors and may not imply a direct causal relationship. Therefore, it is important to interpret the regression analysis results correctly.

3.3.1. SVR

Support Vector Regression (SVR) is a powerful regression method widely accepted in the fields of data analysis and machine learning [56]. SVR is designed to address regression problems and offers several advantages. SVR possesses a flexible framework that can handle different data types and complex relationships effectively [57]. It is capable of addressing both linear and nonlinear regression problems. Its ability to eliminate noisy or outlier data from the dataset contributes to better generalization of the model. Moreover, by employing various kernel functions, SVR can be customized for different regression problems. SVR also provides a set of parameters for fine-tuning the model's complexity and generalization ability, which helps in mitigating issues like overfitting and underfitting.

$$\frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \quad (1)$$

SVR is well-suited for multidimensional data analysis, making it effective for datasets with multiple features and target variables. Particularly, when dealing with datasets containing outlier values, SVR stands out due to its ability to handle such situations [58]. While other regression methods may struggle to tolerate these outlier values, SVR can produce more reliable regression results by disregarding them. Additionally, SVR is an ideal choice for solving regression problems with nonlinear relationships in the dataset. Taking all these advantages into consideration, SVR emerges as a robust and flexible tool for addressing regression problems.

3.3.2. Lasso Regression

Lasso regression is a linear regression method used in regression analysis and also serves as a regularization technique that performs variable selection. It was developed by Robert Tibshirani from Stanford University in 1996 [59]. Unlike traditional linear regression, Lasso adds a penalty term called L1 regularization. The L1 regularization limits the sum of the absolute values of coefficients, indicating that some coefficients approach or become exactly zero. This automatically performs variable selection and identifies significant features in the model. An important parameter of Lasso regression is the regularization parameter, often referred to as "lambda" or "alpha" in Equation 2. This parameter controls the complexity of the regression model; larger values of lambda encourage more variables to approach zero, thus promoting simpler models.

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda \sum_{j=1}^{m}\left|\widehat{\beta}_j\right| \quad (2)$$

Lasso regression offers several advantages. It is particularly useful for high-dimensional datasets as it simplifies the model by performing automatic variable selection. This can reduce data complexity and lead to better generalization. However, Lasso has some disadvantages as well. In cases with a large number of variables, selecting the correct lambda (alpha) value is crucial, as an incorrect choice can negatively impact the model's performance. Lasso may also struggle with handling multicollinearity among independent variables.

In conclusion, Lasso regression is a powerful tool for reducing model complexity and identifying significant features, especially when working with large datasets or high-dimensional data. However, it's essential to carefully select the lambda (alpha) parameter, as it can significantly influence the success of Lasso regression.

3.3.3. Ridge Regression

Ridge regression is a linear regression method and regularization technique used in regression analysis. It began to be widely used in high-dimensional data analysis in the 1970s [60]. Ridge regression builds upon traditional linear regression by adding L2 regularization. L2 regularization restricts the sum of the squares of coefficients, aiming to push some coefficients towards zero while not forcing them to be exactly zero.

Ridge regression is employed to shrink weights and prevent overfitting. This method increases the complexity of the regression model and makes it more resistant to multicollinearity among variables. An important parameter in Ridge regression is the regularization coefficient, often referred to as "lambda" or "alpha." in Equation 3. This parameter controls the complexity of the regression model; larger lambda values push more variables towards zero, encouraging simpler models.

$$L_{ridge}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda \sum_{j=1}^{m} w_j\hat{\beta}_j^2 \quad (3)$$

The advantages of Ridge regression, especially in high-dimensional and collinear data sets, are noteworthy. It can reduce data complexity and enable the model to generalize better. However, it is crucial to choose the correct lambda (alpha) value, as an incorrect choice can negatively impact the model's performance. A drawback of Ridge regression is that L2 regularization pushes coefficients towards zero but not exactly to zero, making it somewhat challenging to obtain a simpler model with zero weights in certain cases. Ridge regression is a valuable tool in data analysis when simplifying the model and preventing overfitting are necessary.

Ridge regression shares some similarities with Lasso Regression, but they are different models. The primary distinction between Ridge regression and Lasso regression lies in the regularization terms used. Ridge regression employs L2 regularization, restricting the sum of the squares of coefficients, while pushing them closer to zero without forcing them to be exactly zero. In contrast, Lasso regression uses L1 regularization, driving some coefficients all the way to zero and automatically performing variable selection. Ridge regression, in contrast to Lasso regression, tends to include multiple variables in the model simultaneously and is more suitable for scenarios where all variables need to be considered. Particularly, when your data contains highly correlated variables, Ridge regression may be a more appropriate choice. Both methods serve as enhanced versions of linear regression to control model complexity and prevent overfitting, but the choice between them depends on the specific requirements of the data and the problem at hand.

3.4. Grey Wolf Optimization Algorithm

Metaheuristic optimization algorithms are powerful approaches widely used in computer science and engineering to solve complex and large-scale optimization problems. These

algorithms are specifically designed to tackle problems for which analytical solutions are difficult or impossible to obtain. One of the prominent advantages of metaheuristic algorithms is their general applicability. In other words, they can be adapted to various types of optimization problems. For instance, when a problem lacks a mathematical model or an analytical solution, metaheuristic algorithms can effectively address such problems.

These algorithms come in various types, including Genetic Algorithms, Simulated Annealing, Particle Swarm Optimization, and Tabu Search, among others. They incorporate heuristics and rules to explore and find the best solutions to problems. However, the use of these algorithms can present some challenges. They do not guarantee an optimal solution and may require different settings for various problems. Additionally, tuning parameters and selecting hyperparameters for the algorithms often demand experience and expertise.

In conclusion, metaheuristic optimization algorithms offer a robust and versatile solution for addressing optimization problems that are too complex or lack analytical solutions. These algorithms find extensive applications in real-world scenarios and scientific research.

Grey Wolf Optimization (GWO) is an effective metaheuristic optimization algorithm inspired by the behaviors of gray wolf packs in nature. This algorithm was proposed by Seyedali Mirjalili , and it is utilized to solve optimization problems, particularly those that are complex and multi-dimensional in nature [59]. The GWO algorithm mimics the natural behavior of gray wolf packs. Each gray wolf represents a potential solution, and these wolves roam the problem space in search of better solutions. The GWO relies on a hierarchical leadership structure among wolves, and they interact with each other to update their leadership positions, and hirearcy of the wolf packs shown in Figure 3.2. This collaborative approach allows wolves to explore the problem space as if stalking their prey and aim to find the best solution.

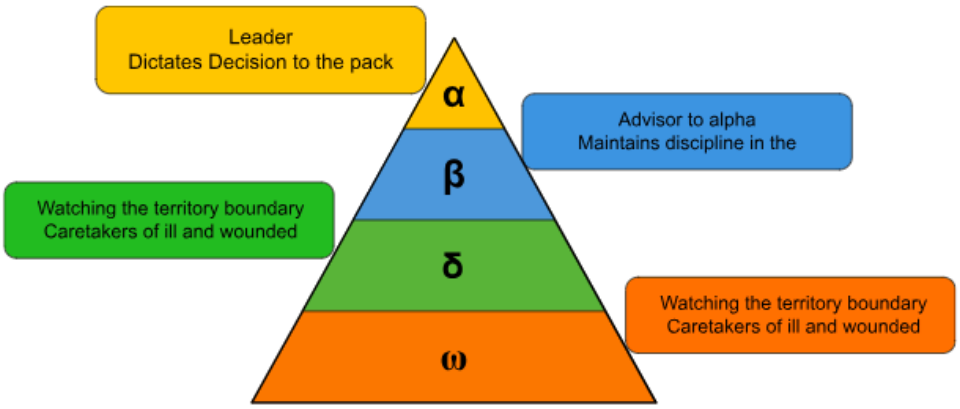

Figure 3.2. Hierarchy of wolf in GWO algorithm.

GWO finds applications in a wide variety of fields, including engineering, artificial intelligence, data mining, electrical engineering, and others [61]. It is especially effective in solving complex and multidimensional optimization problems. Such problems can arise in a variety of practical applications, such as designing artificial neural networks, optimizing supply chain management, planning energy distribution, and financial modelling. Compared to other optimization algorithms, GWO offers distinct advantages It has the potential to further improve its performance as the complexity and size of the problem space increases [38, 39]. Therefore, GWO is considered a powerful tool for researchers and engineers to solve complex problems effectively.

In the first step in the gray wolf algorithm operation, a wolf pack is created. Each wolf represents a potential solution position. With these positions, wolves look for potential solutions to the problem. Within GWO, wolves exist in a hierarchical order. The wolf is called "Alpha", which represents the best solution in the pack. The wolf representing the second-best solution is called "Beta" and the wolf "Delta" represents the third best solution. Other wolves are called "Omega". At the end of each iteration, the wolves move in a certain way. While Alpha, Beta and Delta wolves watch their prey, other wolves try to help these leaders. The wolves' positions are updated, and their new positions are calculated. These new positions represent potential solution positions and are evaluated with the objective function. Finally, the best solution is selected and obtained among Alpha, Beta and Delta wolves. This solution represents the solution to the optimization problem. Gray Wolf

Optimization can be used to solve different problems and is an especially effective tool for complex optimization problems.

The main stages of wolf hunting include tracking, chasing, and approaching the prey; following, surrounding, and harassing the prey until it stops moving; and finally, attacking the prey. Equation 4 and equation 5 are used to mathematically model containment behavior.

$$\vec{D} = \left|\vec{C}.\overrightarrow{X_p}(t) - \vec{X}(t)\right| \tag{4}$$

$$\vec{X}(t+1) = \vec{X}_p(t) - \vec{A}.\vec{D} \tag{5}$$

Here, 't' represents the current iteration, '$\vec{A}$' and '$\vec{C}$' are coefficient vectors, 'X_p' denotes the position vector of the prey, and 'X' represents the position vector of the gray wolf. The vectors Eq. 6 and Eq. 7 are calculated as follows:

$$\vec{A} = 2\vec{a}.\overrightarrow{r_1} - \vec{a} \tag{6}$$

$$\vec{C} = 2.\overrightarrow{r_2} \tag{7}$$

where components of 'a' are linearly decreased from 2 to 0 over the course of iterations, and 'r1' and 'r2' are random vectors in [0, 1] . To update the positions of the wolves optimally, formulas Eq. 8, Eq. 9 and Eq.10 are employed.

$$\overrightarrow{D_\alpha} = |\ \overrightarrow{C_1}.\overrightarrow{X_\alpha} - \vec{X}|\ ,\ \ \overrightarrow{D_\beta} = |\ \overrightarrow{C_2}.\overrightarrow{X_\beta} - \vec{X}|,\ \overrightarrow{D_\delta} = |\ \overrightarrow{C_3}.\overrightarrow{X_\delta} - \vec{X}| \tag{8}$$

$$\overrightarrow{X_1} = \overrightarrow{X_\alpha} - \overrightarrow{A_1}.\left(\overrightarrow{D_\alpha}\right),\ \overrightarrow{X_2} = \overrightarrow{X_\beta} - \overrightarrow{A_2}.\left(\overrightarrow{D_\beta}\right),\ \overrightarrow{X_3} = \overrightarrow{X_\delta} - \overrightarrow{A_3}.\left(\overrightarrow{D_\delta}\right) \tag{9}$$

$$\vec{X}(t+1) = \frac{\overrightarrow{X_1}+\overrightarrow{X_2}+\overrightarrow{X_3}}{3} \tag{10}$$

The 'A' value is randomly selected and controls the behavior of moving towards or away from the prey. 'C' provides random weights to encourage exploration and avoid local optima. The GWO algorithm initially creates a random population, and the wolves then follow the estimated position of the prey. Additionally, the 'A' parameter is gradually reduced to balance exploration and exploitation. The algorithm terminates when the termination criteria are met. The process of the GWO algorithm steps shown in Figure 3.3. and the pseudecode of the algorithm represented in the Figure 3.4.

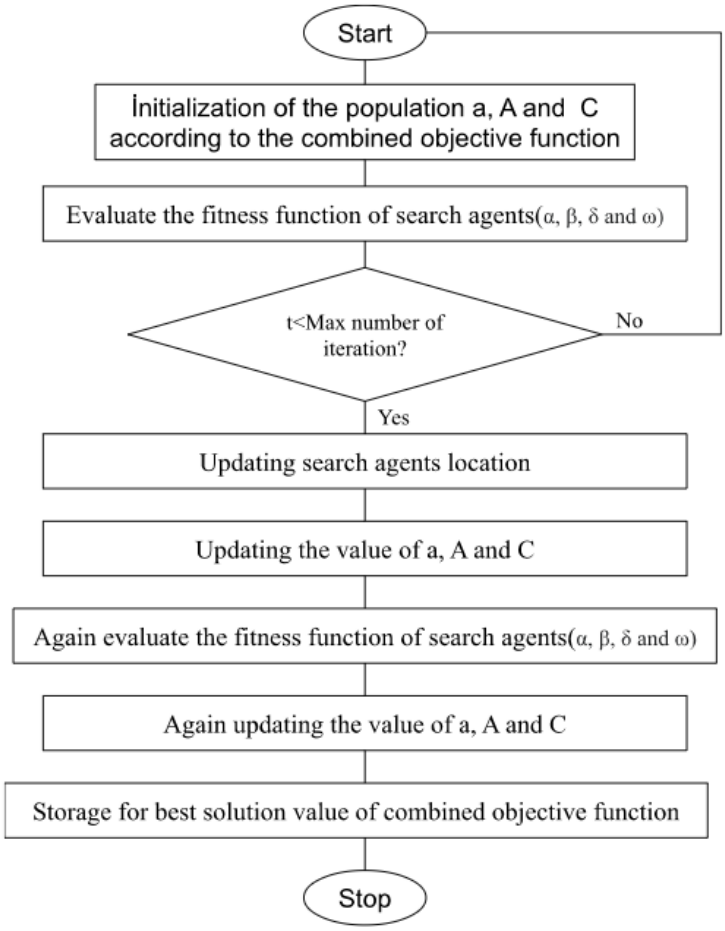

Figure 3.3. Flow Chart of GWO

```
Initialize a population of grey wolves
Initialize alpha, beta, and delta positions randomly
Set maximum iterations or stopping criterion
Set a parameter for exploration rate C
Set a parameter for the initial temperature T (optional, for some variations)
Set a parameter for cooling rate (optional, for some variations)
while (not reached maximum iterations or stopping criterion):
    for each grey wolf:
        Update the fitness value of the current solution
    Sort the grey wolves based on fitness in ascending order
    Calculate the values of a and A:
    a = 2 - 2 * iteration / maximum_iterations
    A = 2 * a * random() - a
    Update the positions of alpha, beta, and delta wolves:
    X_alpha = alpha_position - A * (2 * random() - 1)
    X_beta = beta_position - A * (2 * random() - 1)
    X_delta = delta_position - A * (2 * random() - 1)
    for each grey wolf:
        Update the position of the wolf based on alpha, beta, and delta positions:
        X_i = (X_alpha + X_beta + X_delta) / 3
        If X_i is outside the search space boundaries:
            Bring X_i back within the boundaries
        Apply a random exploration factor:
        X_i = X_i + C * (2 * random() - 1)
        Update the fitness value of the new solution
        If the new solution is better than the previous:
            Replace the old solution with the new one
    Decrease the temperature T (optional, for some variations)
Return the best solution found
```

Figure 3.4. Pseudo Code of the GWO Algorithm

In conclusion, Grey Wolf Optimization is a meta-heuristic optimization algorithm inspired by natural behavior. By mimicking these behaviors, it solves complex optimization problems and is widely applied in industrial and scientific fields. This approach streamlines and optimizes the problem-solving process by effectively dealing with complexity.

3.5. Evaluation Metrics

Evaluation metrics are criteria or criteria used to objectively measure and evaluate the performance of a system, product or process. Evaluation metrics can help determine the degree to which certain goals or objectives are achieved and are often expressed in quantitative data. We used MAE, MSE, RMSE and R^2 evaluation metrics in our study.

3.5.1. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is a metric that measures how much forecasts deviate from actual values. To calculate the MAE, the absolute difference of each estimate from the true value is taken, these differences are summed, and finally, the total differences are divided by

the number of data points. The MAE gives an average measure of the absolute values of the prediction errors.

$$MAE = \frac{1}{n}\sum |y - \hat{y}| \quad (11)$$

Here:

n represents the number of data points.

y implies the real values.

ŷ implies the estimated values.

|...| sign represents the absolute value.

3.5.2. Mean Squared Error (MSE)

Mean Squared Error (MSE) is a metric that measures how much the estimates deviate from the true values. The MSE is calculated by squaring the difference of each estimate from the true value, summing these differences, and finally dividing the total differences by the number of data points. The MSE gives an average measure of the squares of the prediction errors.

$$MSE = \frac{1}{n}\sum (y - \hat{y})^2 \quad (12)$$

Here:

n is the number of data points.

y are real values.

ŷ are the estimated values.

The ^ sign denotes its exponent.

3.5.3. Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is a metric obtained by taking the square root of the MSE. The RMSE returns the square root of the mean of the squares of the prediction errors. The RMSE is a metric that measures how much the estimates deviate from the true values and is expressed in the same unit as the MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum (y - \hat{y})^2} \quad (13)$$

3.5.4. Coefficient of Determination (R^2)

Coefficient of determination (R^2) is a metric that measures how well the linear regression model fits. R^2 expresses the ratio of the variance in the dependent variable to the variance explained by the independent variables. The R^2 value takes a value between 0 and 1, the closer to 1 the better the model explains the data. If the R^2 value is 0, it means that the model cannot explain the data at all.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{14}$$

$$SST = \sum (y - \hat{y})^2 \tag{15}$$

$$SSR = \sum (y' - \hat{y}')^2 \tag{16}$$

$$SSE = \sum (y - y')^2 \tag{17}$$

Here:

SSR is the sum of the squares of the deviations of the predicted values from the true values. SST refers to the sum of the squares of the deviations of the true values from the mean value. SSE represents to sum of Squares Error.

4. RESULTS AND DISCUSSION

In this section, the steps taken in the study are mentioned in detail and the results are given. In SVR, LASSO, and Ridge regressions, parameters of the regression models were optimized to enhance the performance of the regression models and reduce the problem of overfitting. The optimized model shows improved performance, controls the issue of overfitting, and enhances the generalization ability, making the models more adaptable and better suited to different datasets.

4.1. Experiments

Without applying any optimization algorithms tests were conducted using SVR, Lasso, and Ridge algorithms to assess the performance of the models. During these processes, SVR (Support Vector Regression), Lasso, and Ridge regression algorithms were used. The parameters of the algorithms used are summarized below:

- SVR: degree=3, C=1.0, gamma=scale, kernel=rbf,
- Lasso: alpha=1.0, max_iter=1000, selection= cyclic
- Ridge: alpha=1.0, solver=auto

Table 4.1 summarizes the results of the regression algorithms applied after preprocessing of the data set and imputation of missing data values. To evaluate the results, MAE, MSE, RMSE, and R^2 values were used. After data preprocessing and missing data imputation, as seen in Table 4.1, the results were found to be quite satisfactory when the Lasso regression model was applied compared to other regression models. As a result, when the Lasso regression model was applied to the dataset, it yielded very successful results according to the MAE, MSE, RMSE, and R^2 metrics, with values of 0.038, 0.006, 0.078, and 0.973, respectively. These results indicate that the model fits the data well and makes accurate predictions.

Table 4.1. Comparison of regression results according to evaluation metrics

METRICS	SVR	**LASSO**	RİDGE
MAE	0.159	**0.038**	0.073
MSE	0.027	**0.006**	0.089
RMSE	0.166	**0.078**	0.299
R^2	0.881	**0.973**	0.617

In this study, various regression models were applied to a dataset, and the results were evaluated. Particularly noteworthy were the results obtained by the Lasso regression model in predicting the target variable in the dataset.

The parameters of the regression models, SVR, LASSO, and Ridge, were optimized to enhance the performance of the regression models and mitigate the issue of overfitting. The Grey Wolf algorithm is used to optimize alpha, beta and C values of the ML algorithms. Applying GWO to algorithm the GWO parameters selected as;

SearchAgents_no=20, T=100, dim=2, lb=0.01, ub=10

"SearchAgents_no," represents the number of wolves or search agents, and "T," denotes the maximum number of iterations. "dim" indicates the number of variables to be optimized, while "lb" (lower bound) and "ub" (upper bound) specify the acceptable range of values for the variables under optimization. These parameters are configured based on the characteristics of the task to be optimized and the setup of the algorithm.

To enhance the performance of the SVR model, two crucial hyperparameters are considered: C and gamma. The C parameter controls the regularization coefficient, where low C values result in more regularization and a less complex model, while high C values provide less regularization and a more complex model. The gamma parameter determines the width of the RBF kernel, with small gamma values creating models more tailored to the training data, and large gamma values enhancing generalization ability. Optimizing C and gamma values using the GWO algorithm ensures the model's resilience against overfitting and better adaptation to different datasets.

When optimizing Lasso regression using GWO, the obtained optimal alpha parameter may approach a low value. In Lasso regression, alpha controls the regularization coefficient and encourages regression coefficients to approach zero. In other words, low alpha values pull more regression coefficients towards zero, resulting in a sparser model. The goal of GWO is to find the optimal alpha value that best fits the dataset while controlling the complexity of the model. This way, the model becomes more resistant to overfitting and more generalizable. Finding the optimal alpha parameter enhances the model's performance by promoting accurate learning and preventing unnecessary complexity.

GWO algorithm derives optimal alpha parameter tends to converge towards higher values for Ridge Regression. In Ridge regression, the alpha parameter implies the regularization coefficient, organizing a gradual approach of regression coefficients towards zero without enforcing absolute zeroing. Elevated alpha values induce regression coefficients to approximate zero, distinguishing itself from Lasso regression where a strict zeroing effect is observed. The alpha parameter in Ridge regression, constituting the L2 regularization term, meticulously manages the magnitudes of regression coefficients. This strategic regularization mitigates the peril of overfitting and augments the model's capacity for generalization. Ridge regression, fundamentally, endeavors to strike a harmonious equilibrium between fostering precise learning and tempering the intricacies of the model. Through the application of the GWO algorithm, the optimization process fine-tunes the alpha parameter of Ridge regression, culminating in a regression model characterized by heightened reliability and enhanced generalizability.

For SVR, the optimal 'C' and gamma values were found to be 0.072 and 0.107 respectively. The optimal 'alpha' value was determined to be 0.011 for Lasso regression. The best 'alpha' value was found to be 0.013 in the case of Ridge regression. As a result of Hyperparameter Optimization is comparison of the regression models are presented in Table 4.2. When compared with other regression models, the Lasso regression model achieved the best results. The results are as follows for MAE, MSE, RMSE, and R^2 values are 0.030, 0.003, 0.062, and 0.983 respectively.

Table 4.2. Comparison of GWO-optimized regression results based on evaluation metrics

METRICS	SVR	**LASSO**	RİDGE
MAE	0.145	**0.030**	0.063
MSE	0.024	**0.003**	0.079
RMSE	0.152	**0.062**	0.289
R^2	0.890	**0.983**	0.623

When the Grey Wolf Optimizer (GWO) was applied, a significant improvement in the performance of the Lasso regression model was observed. Therefore, the results of this study demonstrate that Lasso regression, both in its traditional form and when optimized, is an effective method for solving regression problems in the dataset.

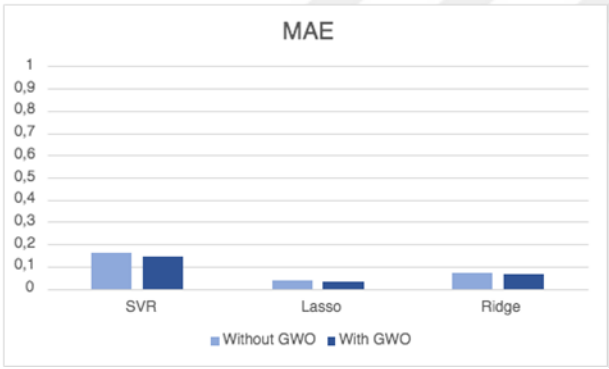

Figure 4.1. Graph of MAE values of models

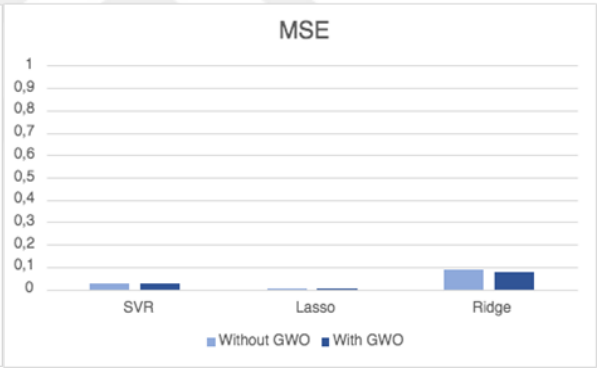

Figure 4.2. Graph of MSE values of models

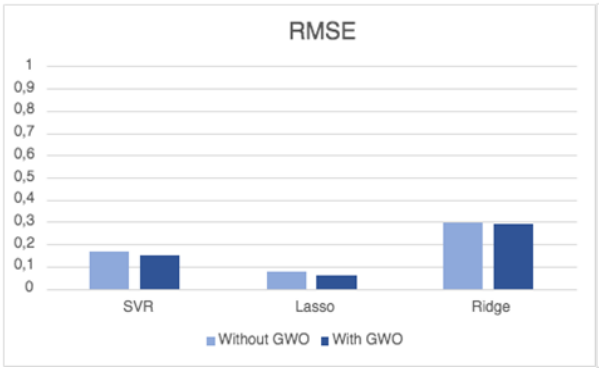

Figure 4.3. Graph of RMSE valuesof models

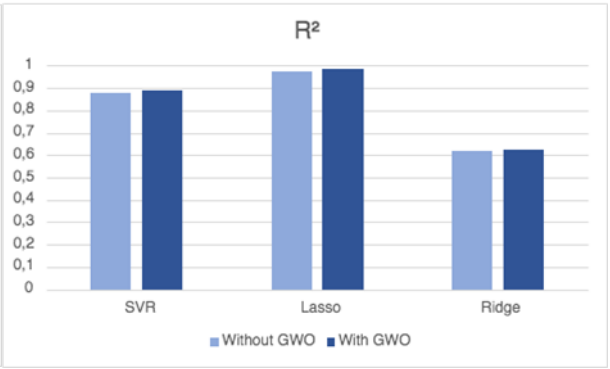

Figure 4.4. Graph of R^2 values of models

Figure 4.1, Figure 4.2, Figure 4.3 and Figure 4.4, present a comparative display of the results obtained for the SVR, Lasso, and Ridge regression models before and after the application of GWO. When examining the results, it is evident that there is an improvement in every evaluation metric after the application of GWO. However, Lasso regression consistently provided the best results.

Figure 4.1, Figure 4.2, Figure 4.3 and Figure 4.4, shows a comparative graph of the results before and after the application of the Grey Wolf Optimization technique to machine learning algorithms. These results indicate that the model effectively describes the dataset and has a high predictive ability. In particular, the low MAE and MSE values demonstrate how close the model's predictions are to the actual values. The high R^2 value also emphasizes the model's ability to explain the variance of the independent variables in the target variable.

In conclusion, the Lasso regression model optimized with GWO emerges as a successful option for predicting the target variable in the dataset.

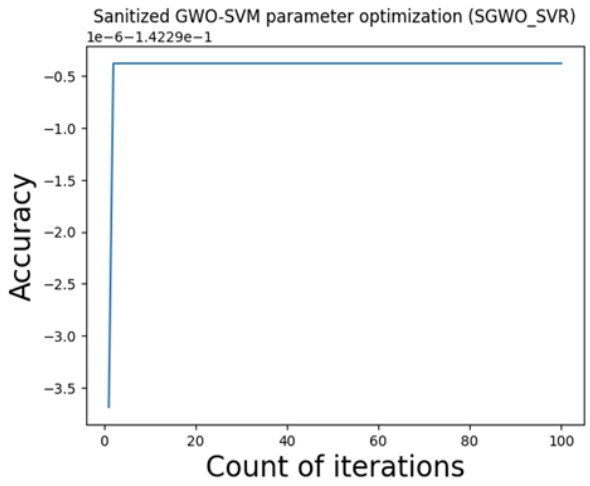

Figure 4.5. Accuracy chart of SVR model using GWO

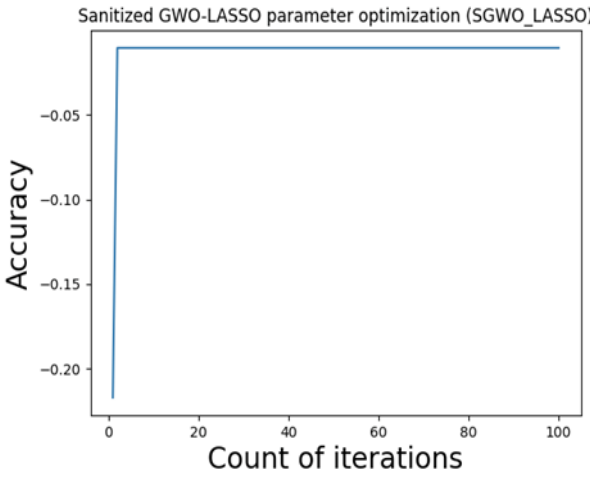

Figure 4.6. Accuracy chart of Lasso model using GWO

Figure 4.5 and Figure 4.6 shows the Gray Wolf algorithm significantly increases the accuracy value in both models. Looking at the value reached after the 5th iteration, it can be seen that it is a fast-converging algorithm.

4.2. Discussion

This thesis aims to forecast B12 vitamin levels based on the findings of blood tests. First off, the dataset's erroneous gender and date columns have been eliminated. Too much blank data is present in some columns. Additionally, removed from the database were rows with more than 80% blank data. Iterative imputation was used to fill in the gaps left by missing data in the context of this thesis. The use of Grey Wolf Optimization has been made. Lasso regression was also used for estimation. Then, the values of MAE, MSE, RMSE, and R^2 were computed.

On Colab, all work was completed. The dataset was first added, and it experienced some changes. For instance, changes have been made so that columns with more than 20% null values are deleted. Imputation techniques were then used.

The best imputation method was iterative. Although feature selection was used, it was unsuccessful. After that, regression was used for estimation. Lasso regression produced the best results. Following that, GWO was used. The results showed that, once GWO was used, the Lasso regression's results improved.

5. CONCLUSIONS

This thesis aims to predict B12 vitamin levels based on blood test results. In the first step, inappropriate columns in the dataset, such as gender and date, were removed. Additionally, 40 rows of data with missing values for B12 vitamin predictions were excluded. Rows with more than 80% missing data in the data table were also eliminated. Iterative imputation was used to handle missing data in this study. Furthermore, SVR, Lasso, and Ridge regression models were employed for predictions. Subsequently, MAE, MSE, RMSE, and R^2 values were calculated to evaluate the performance of the predictions. Enhancing the performance of the regression models and reducing the overfitting in SVR, LASSO, and Ridge regressions, the parameters of the regression models were optimized. The optimized model shows improved performance, controls the issue of overfitting, and enhances the generalization ability, making the models more adaptable and better suited to different datasets. For SVR, the optimal 'C' and 'gamma' values were found to be 0.072 and 0.107, respectively. For Lasso regression, the optimal 'alpha' value was determined to be 0.011. In the case of Ridge regression, the best 'alpha' value was found to be 0.013. Remarkably, the Lasso regression model consistently yielded the best results in both stages. The R^2 evaluation metric showed that the Lasso Regression result was 0.973, which increased to 0.983 after optimizing with the Lasso Regression Grey Wolf algorithm. The application of these techniques significantly improved the thesis results. Moreover, when compared to similar studies, this model's performance was found to be higher and superior.

All research steps were carried out on the Colab platform. Then, regression models were used for the prediction process, with the Lasso regression model providing the most effective results. Subsequently, the Grey Wolf algorithm was introduced to further enhance the model's performance. The results demonstrated that the Grey Wolf algorithm improved the performance of Lasso regression.

The results were evaluated through four important metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 (Coefficient of Determination). These metrics are crucial in indicating how close the model's predictions are to the actual values.

When GWO was used, the results showed a significant improvement in all four metrics. This suggests that the Grey Wolf algorithm substantially enhanced the model's prediction

performance. These results underscore the effectiveness of GWO as a strategy that assists the model in making more accurate predictions.

This study highlights the importance of blood analysis in medicine and the power of regression analysis. Particularly, a high R^2 value indicates that the model explains a significant portion of the variance in the dependent variable and that the predictions are very close to the actual values. This can increase the usability of the model in clinical applications, enabling healthcare professionals to predict blood analyses more accurately.

In conclusion, this study emphasizes the potential of the Grey Wolf Optimization (GWO) method for improving the performance of regression models. In future research, it is believed that results can be further enhanced by using more optimization techniques and various regression methods. Additionally, the dissemination and application of these methods in different datasets and medical fields will be a critical step for future research.

6. FUTURE WORKS

There are several potential opportunities for future research. First and foremost, going beyond the regression models used in the current study and exploring different types of models could enhance prediction performance. Evaluating the impact of more complex methods, such as deep learning models, is important in this context. Furthermore, expanding the dataset beyond data obtained from a single region and accessing larger and more diverse data sources can improve the model's generalization ability. It's also important to consider real-time updates to keep your dataset up to date. Additionally, developing the numerical dataset in a way that it can be categorized and used for classification can broaden the range of data analysis. Finally, adapting the success of this study to other health predictions or different application domains can enable these techniques to have a broader impact.

REFERENCES

[1] S. Kuriyama, "Interpreting the history of bloodletting," *Journal of the History of Medicine and Allied Sciences*, vol. 50, no. 1, pp. 11–46, 1995.

[2] A. Mudry and J. R. Young, "A Critical Evaluation of 'The Ear that Hears Badly' in the Ebers Papyrus," *Otology & Neurotology*, vol. 42, no. 8, pp. 1285–1290, 2021.

[3] M. Meyerhof, "Ibn An-Nafîs (XIIIth Cent.) and His Theory of the Lesser Circulation," Isis, vol. 23, no. 1, pp. 100–120, Jun. 1935, doi: 10.1086/346938.

[4] P. Distelzweig, "'Mechanics' and Mechanism in William Harvey's Anatomy: Varieties and Limits," History, Philosophy and Theory of the Life Sciences, vol. 14, pp. 117–140, 2016, doi: 10.1007/978-94-017-7353-9_6.

[5] "Classic Papers in Critical Care" Classic Papers in Critical Care, 2008, doi: 10.1007/978-1-84800-145-9

[6] K. Pelis, "Review of Blood: An Epic History of Medicine and Commerce," *Isis*, vol. 94, no. 2, pp. 355–356, 2003, doi: 10.1086/379417.

[7] J. Blundell, "Experiments on the transfusion of blood by the syringe," *Medico-chirurgical transactions*, vol. 9, no. Pt 1, p. 56, 1818.

[8] R. Baş, "Türkiye'de Kan Nakli (Transfüzyon) Tarihçesi (1920-1960)," *2. ERASMUS Uluslararası Akademik Araştırmalar Sempozyumu*, p. 163, 2019.

[9] W. Apollon et al., "A Beginner's Guide to Different Types of Microscopes," Microscopic Techniques for the Non-Expert, pp. 1–23, 2022, doi: 10.1007/978-3-030- 99542-3_1.

[10] J. J. M. L. Hoffmann and E. Urrechaga, "Additions to the paper 'The Journey of Laboratory Medicine' from a broader perspective," J Clin Pathol, Apr. 2023, doi: 10.1136/JCP-2023-208884.

[11] R. Guthrie, "Blood Screening for Phenylketonuria," JAMA, vol. 178, no. 8, pp. 863–863, Nov. 1961, doi: 10.1001/JAMA.1961.03040470079019.

[12] L. Erdim, N. Zengin, and B. Karaca, “Investigation of the preanalytical process practices in primary care in Istanbul regarding the newborn screening tests,” *Turkish Journal of Biochemistry*, vol. 48, no. 1, pp. 12–18, Mar. 2023, doi: 10.1515/tjb-2022-0116.

[13] M. Napiórkowska-Orkisz, A. Gutysz-Wojnicka, M. Tanajewska, and I. Sadowska-Krawczenko, “Evaluation of methods to minimize pain in newborns during capillary blood sampling for screening: a randomized clinical trial,” *International Journal of Environmental Research and Public Health*, vol. 19, no. 2, p. 870, 2022.

[14] E. A. Cooper and C. Funk, “EXPERIMENTS ON THE CAUSATION OF BERI-BERI.:(PRELIMINARY COMMUNICATION.),” *The Lancet*, vol. 178, no. 4601, pp. 1266–1267, 1911.

[15] C. Funk, “ii. The Substance from Yeast and certain Foodstuffs which prevents Polyneuritis (Beri-Beri).,” Br Med J, 1912.

[16] S. Buesing, M. Costa, J. M. Schilling, and T. Moeller-Bertram “Narrative Review Vitamin B12 as a Treatment for Pain”, Accessed: Jun. 06, 2023.

[17] A. Fanidi et al., “Is high vitamin B12 status a cause of lung cancer?” Int J Cancer, vol. 145, no. 6, pp. 1499–1503, Sep. 2019, doi: 10.1002/IJC.32033.

[18] B. H. R. Wolffenbuttel, H. J. C. M. Wouters, M. R. Heiner-Fokkema, and M. M. van der Klauw, “The Many Faces of Cobalamin (Vitamin B12) Deficiency,” *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*, vol. 3, no. 2, pp. 200–214, Jun. 2019, doi: 10.1016/j.mayocpiqo.2019.03.002.

[19]“Gelişmiş ülkelerde Alzheimer hastalığı azalırken, Türkiye’de artıyor! | Türkiye Alzheimer Derneği.” Accessed: Nov. 09, 2023. [Online]. Available: https://www.alzheimerdernegi.org.tr/gelismis-ulkelerde-alzheimer-hastaligi-azalirken-turkiyede-artiyor/

[20] A. A. Lauer et al., “Mechanistic Link between Vitamin B12 and Alzheimer’s Disease,” Biomolecules 2022, Vol. 12, Page 129, vol. 12, no. 1, p. 129, Jan. 2022, doi: 10.3390/BIOM12010129.

[21] X. Chen et al., “Non-invasive early detection of cancer four years before conventional diagnosis using a blood test,” Nature Communications 2020 11:1, vol. 11, no. 1, pp. 1– 10, Jul. 2020, doi: 10.1038/s41467-020-17316-z.

[22] D. Monett, C. W. P. Lewis, and K. R. Thórisson, “Journal of Artificial General Intelligence Special Issue ‘On Defining Artificial Intelligence’-Commentaries and Author’s Response”, doi: 10.2478/jagi-2020-0003.

[23] S. P. Somashekhar *et al.*, “Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board,” *Annals of Oncology*, vol. 29, no. 2, pp. 418–423, 2018.

[24] I. M. Nasser and S. S. Abu-Naser, “Lung cancer detection using artificial neural network,” *International Journal of Engineering and Information Systems (IJEAIS)*, vol. 3, no. 3, pp. 17–23, 2019.

[25] H.-Y. ; Chiu, H.-S. ; Chao, F. Petrella, H.-Y. Chiu, H.-S. Chao, and Y.-M. Chen, “Application of Artificial Intelligence in Lung Cancer,” Cancers 2022, Vol. 14, Page 1370, vol. 14, no. 6, p. 1370, Mar. 2022, doi: 10.3390/CANCERS14061370.

[26] I. M. Nasser and S. S. Abu-Naser, “Lung Cancer Detection Using Artificial Neural Network.” Rochester, NY, Mar. 01, 2019. Accessed: Nov. 09, 2023. [Online]. Available: https://papers.ssrn.com/abstract=3369062

[27] C. Guo and H. Li, “Application of 5G network combined with AI robots in personalized nursing in China: A literature review,” Front Public Health, vol. 10, Aug. 2022, doi: 10.3389/FPUBH.2022.948303/PDF.

[28] G. M. Minopoulos, V. A. Memos, K. D. Stergiou, C. L. Stergiou, and K. E. Psannis, “A Medical Image Visualization Technique Assisted with AI-Based Haptic Feedback for Robotic Surgery and Healthcare,” Applied Sciences 2023, Vol. 13, Page 3592, vol. 13, no. 6, p. 3592, Mar. 2023, doi: 10.3390/APP13063592.

[29] E. I. George, T. C. Brand, A. LaPorta, J. Marescaux, and R. M. Satava, “Origins of Robotic Surgery: From Skepticism to Standard of Care,” JSLS, vol. 22, no. 4, Oct. 2018, doi: 10.4293/JSLS.2018.00039.

[30] A. L. G. Morrell et al., “The history of robotic surgery and its evolution: When illusion becomes reality,” Rev Col Bras Cir, vol. 48, pp. 1–9, 2021, doi: 10.1590/0100-6991e20202798.

[31] Z. Xu and Y. Zhang, “What’s new in artificially intelligent joint surgery in China? The minutes of the 2021 IEEE ICRA and literature review,” Arthroplasty, vol. 4, no. 1, pp. 1–11, Dec. 2022, doi: 10.1186/S42836-021-00109-0/TABLES/1.

[32] A. Agrawal, J. S. Gans, and A. Goldfarb, “Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction,” Journal of Economic Perspectives, vol. 33, no. 2, pp. 31–50, Mar. 2019, doi: 10.1257/JEP.33.2.31.

[33] M. Akhtar and S. Moridpour, “A Review of Traffic Congestion Prediction Using Artificial Intelligence,” J Adv Transp, vol. 2021, 2021, doi: 10.1155/2021/8878011.

[34] Z.-H. Zhou, *Machine Learning*. Springer Nature, 2021.

[35] “Data Mining Techniques: Review,” International Journal of Data Science Research.

[36] P. Kumar and A. Gupta, “Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey,” J Comput Sci Technol, vol. 35, no. 4, pp. 913– 945, Jul. 2020, doi: 10.1007/S11390-020-9487-4.

[37] D. Kirk, C. Catal, and B. Tekinerdogan, “Predicting Plasma Vitamin C Using Machine Learning,” Applied Artificial Intelligence, vol. 36, no. 1, Dec. 2022, doi: 10.1080/08839514.2022.2042924.

[38] S. KILIÇARSLAN, “PSO + GWO: a hybrid particle swarm optimization and Grey Wolf optimization based Algorithm for fine-tuning hyper-parameters of convolutional neural networks for Cardiovascular Disease Detection,” *J Ambient Intell Human Comput*, vol. 14, no. 1, pp. 87–97, Jan. 2023, doi: 10.1007/s12652-022-04433-4.

[39] R. Mohakud and R. Dash, “Designing a grey wolf optimization based hyper-parameter optimized convolutional neural network classifier for skin cancer detection,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, Part B, pp. 6280–6291, Sep. 2022, doi: 10.1016/j.jksuci.2021.05.012.

[40] R. kun Zhang, Q. Xiao, S. lang Zhu, H. yan Lin, and M. Tang, “Using different machine learning models to classify patients into mild and severe cases of COVID-19 based on multivariate blood testing,” J Med Virol, vol. 94, no. 1, pp. 357–365, Jan. 2022, doi: 10.1002/JMV.27352.

[41] V. Formica et al., “Complete blood count might help to identify subjects with high probability of testing positive to SARS-CoV-2,” Clinical Medicine, vol. 20, no. 4, pp. e114–e119, Jul. 2020, doi: 10.7861/CLINMED.2020-0373.

[42] A. Banerjee *et al.*, “Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population,” *International immunopharmacology*, vol. 86, p. 106705, 2020.

[43] E. Avila, A. Kahmann, C. Alho, and M. Dorn, “Hemogram data as a tool for decision-making in COVID-19 management: applications to resource scarcity scenarios,” *PeerJ*, vol. 8, p. e9482, 2020.

[44] R. P. Joshi et al., “A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results,” Journal of Clinical Virology, vol. 129, p. 104502, Aug. 2020, doi: 10.1016/J.JCV.2020.104502.

[45] S. Kurstjens, T. De Bel, A. Van Der Horst, R. Kusters, J. Krabbe, and J. Van Balveren, “Automated prediction of low ferritin concentrations using a machine learning algorithm,” *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 60, no. 12, pp. 1921–1928, Nov. 2022, doi: 10.1515/cclm-2021-1194.

[46] Y. Pan et al., “Machine Learning Prediction of Iron Deficiency Anemia in Chinese Premenopausal Women 12 Months after Sleeve Gastrectomy,” Nutrients 2023, Vol. 15, Page 3385, vol. 15, no. 15, p. 3385, Jul. 2023, doi: 10.3390/NU15153385.

[47] N. James, L. Gerrish, N. Rokotyan, and P. A. Gladding, “Machine Learning Applied to Routine Blood Tests and Clinical Metadata to Identify and Classify Heart failure”, doi: 10.1101/2021.07.26.21261115.

[48] G. Gunčar et al., “An application of machine learning to haematological diagnosis,” Scientific Reports 2018 8:1, vol. 8, no. 1, pp. 1–12, Jan. 2018, doi: 10.1038/s41598-017-18564-8.

[49] C. E. Valderrama, D. J. Niven, H. T. Stelfox, and J. Lee, “Predicting abnormal laboratory blood test results in the intensive care unit using novel features based on information theory and historical conditional probability: observational study,” *JMIR Medical Informatics*, vol. 10, no. 6, p. e35250, 2022.

[50] N. Meng, P. Zhang, J. Li, J. He, and J. Zhu, “Prediction of Coronary Heart Disease Using Routine Blood Tests.” arXiv, Sep. 11, 2018. Accessed: Nov. 09, 2023. [Online]. Available: http://arxiv.org/abs/1809.09553

[51] H. Tamune, J. Ukita, Y. Hamamoto, H. Tanaka, K. Narushima, and N. Yamamoto, “Efficient Prediction of Vitamin B Deficiencies via Machine-Learning Using Routine Blood Test Results in Patients With Intense Psychiatric Episode,” Front Psychiatry, vol. 10, p. 491435, Feb. 2020, doi: 10.3389/FPSYT.2019.01029/BIBTEX.

[52] S. S. Sharifmousavi and M. S. Borhani, “Support vectors machine-based model for diagnosis of multiple sclerosis using the plasma levels of selenium, vitamin B12, and vitamin D3,” Inform Med Unlocked, vol. 20, p. 100382, Jan. 2020, doi: 10.1016/J.IMU.2020.100382.

[53] W. C. Lin and C. F. Tsai, “Missing value imputation: a review and analysis of the literature (2006–2017),” Artif Intell Rev, vol. 53, no. 2, pp. 1487–1509, Feb. 2020, doi: 10.1007/S10462-019-09709-4/METRICS.

[54] M. Song et al., “A Review of Integrative Imputation for Multi-Omics Datasets,” Front Genet, vol. 11, p. 570255, Oct. 2020, doi: 10.3389/FGENE.2020.570255/BIBTEX.

[55] P. C. Austin, I. R. White, D. S. Lee, and S. van Buuren, “Missing Data in Clinical Research: A Tutorial on Multiple Imputation,” Canadian Journal of Cardiology, vol. 37, no. 9, pp. 1322–1331, Sep. 2021, doi: 10.1016/J.CJCA.2020.11.010.

[56] L. Kunhui, L. Qiang, Z. Changle, and Y. Junfeng, “Time series prediction based on linear regression and SVR,” Proceedings - Third International Conference on Natural Computation, ICNC 2007, vol. 1, pp. 688–691, 2007, doi: 10.1109/ICNC.2007.780.

[57] B. Zhang, H. Ren, G. Huang, Y. Cheng, and C. Hu, “Predicting blood pressure from physiological index data using the SVR algorithm,” BMC Bioinformatics, vol. 20, no. 1, pp. 1–15, Feb. 2019, doi: 10.1186/S12859-019-2667-Y/FIGURES/9.

[58] M. H. Zangooei, J. Habibi, and R. Alizadehsani, “Disease Diagnosis with a hybrid method SVR using NSGA-II,” Neurocomputing, vol. 136, pp. 14–29, Jul. 2014, doi: 10.1016/J.NEUCOM.2014.01.042.

[59] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[60] A. E. Hoerl and R. W. Kennard, “Ridge Regression: Applications to Nonorthogonal Problems,” Technometrics, vol. 12, no. 1, pp. 69–82, 1970, doi: 10.1080/00401706.1970.10488635.

[61] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey Wolf Optimizer,” Advances in Engineering Software, vol. 69, pp. 46–61, Mar. 2014, doi: 10.1016/J.ADVENGSOFT.2013.12.007.