Heliyon 9 (2023) e15940

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Artificial intelligence to assist specialists in the detection of haematological diseases

Sergio Diaz-del-Pino [a,*], Roberto Trelles-Martinez [b], F.A. González-Fernández [c], Nicolas Guil [a]

[a] Computer Architecture Department, University of Malaga, Spain
[b] Hematology and Hemotherapy Service, Fundación Alcorcón University Hospital (Madrid), Spain
[c] Hematology and Hemotherapy Service, Clínico San Carlos University Hospital (Madrid), Spain

ARTICLE INFO

Keywords:
Aiding clinicians
Hemograms
Machine learning
Classification
Artificial intelligence
Assist
Neural network
Anaemia
Complete blood count
Haematology
Diagnosis

ABSTRACT

Artificial intelligence, particularly the growth of neural network research and development, has become an invaluable tool for data analysis, offering unrivalled solutions for image generation, natural language processing, and personalised suggestions. In the meantime, biomedicine has been presented as one of the pressing challenges of the 21st century. The inversion of the age pyramid, the increase in longevity, and the negative environment due to pollution and bad habits of the population have led to a necessity of research in the methodologies that can help to mitigate and fight against these changes.

The combination of both fields has already achieved remarkable results in drug discovery, cancer prediction or gene activation. However, challenges such as data labelling, architecture improvements, interpretability of the models and translational implementation of the proposals still remain. In haematology, conventional protocols follow a stepwise approach that includes several tests and doctor-patient interactions to make a diagnosis. This procedure results in significant costs and workload for hospitals.

In this paper, we present an artificial intelligence model based on neural networks to support practitioners in the identification of different haematological diseases using only rutinary and inexpensive blood count tests. In particular, we present both binary and multiclass classification of haematological diseases using a specialised neural network architecture where data is studied and combined along it, taking into account the clinical knowledge of the problem, obtaining results up to 96% accuracy for the binary classification experiment. Furthermore, we compare this method against traditional machine learning algorithms such as gradient boosting decision trees and transformers for tabular data. The use of these machine learning techniques could reduce the cost and decision time and improve the quality of life for both specialists and patients while producing more precise diagnoses.

1. Introduction

In the last few years, the information generated by humans has been growing continuously, along with the use of technology in our

* Corresponding author.
E-mail address: sergiodiazdp@uma.es (S. Diaz-del-Pino).

https://doi.org/10.1016/j.heliyon.2023.e15940
Received 24 January 2023; Received in revised form 27 April 2023; Accepted 28 April 2023
Available online 3 May 2023
2405-8440/© 2023 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

day-to-day lives. The increase in computational resources, the expansion of the internet and, lately, smartphones, the social networks and the use of wearable devices, are contributing to an increase in the available amount of data which demands to be analysed in order to take advantage of it. To put it into perspective, we created 2.5 quintillion data bytes daily in 2020, including more than 65 billion Whatsapp messages per day and 3.5 billion Google searches. This has resulted in a growth rate of 2x every 40 months [1].

For this purpose, Artificial Intelligence (AI) has become one of the most interesting fields for analysing huge amounts of data. Particularly, neural networks are already making a difference in solving diverse problems, such as human gait recognition [2,3], self-driving cars [4] or text-related tasks [5]. Almost every big company in the world is working in this field making them an indispensable part of our daily life. In biomedicine, there are also several examples where AI is providing innovative solutions for problems such as drug discovery [6], gene expression [7] or cancer prediction [8].

Biomedicine doesn't intend to fall behind this trend [9,10] and the predictions already point towards the field with the highest data growth for 2025 [11]. This mainly includes three different types: 1) genomic data (i.e. gen banks, genomes and short sequences), 2) biometric, which includes information collected by body sensors such as insulin or heart rate and 3) clinical data, which includes from clinical history, analysis, X-ray images, etc. With all this new information, the way experiments are being performed has changed from small to big populations, which also generates more results that need to be analysed from a holistic perspective.

In particular, haematology, which is responsible for the study and diagnosis of diseases related to blood, presents some interesting opportunities for AI [12,13]. The workflow for making a diagnosis starts with a blood test called Hemogram or Complete Blood Count. This test consists of the extraction of a single blood vial for the measurement of several blood features such as haemoglobin or leukocytes, among others. A hemogram is the cheapest test in the protocol and is usually done on a regular basis in all medical evaluations, but further testing [14] (i.e. Peripheral blood smear, genetic test) is needed to make a precise diagnosis, which entails increased human resources with estimated costs of up to €2000 per patient [15,16].

We can find several attempts for an intelligent classification of haematological disease by using blood features in the literature. There is a vast majority of binary classification approaches between healthy patients and a specific disease or between two specific diseases [17–20] that have been developed using more traditional machine learning algorithms such as Support Vector Machines [21], which search for optimal hyperplanes to separate classes in a defined space; Decision Trees [22], which creates trees where each node uses a feature and each branch represents a decision based on that feature; or Multi-Layer Perceptrons [23], which are standard fully connected feed-forward neural networks. There are only a few examples where multiclass classification is used [24] and fewer when it is combined with thousands instead of hundreds of samples [25]. Additionally, Machine Learning algorithms are usually treated as a black box in the majority of scientific papers, and regarding biomedicine, it is especially important to understand that the reasons for our classifications are based on scientific knowledge [26].

In this manuscript, we present an artificial intelligence method for classifying haematological diseases. Specifically, we have developed a neural network architecture based on the traditional protocol used by specialists to diagnose haematological diseases associated with anaemia. This neural network, which uses hemograms as data input and combines specific features throughout the architecture, is able to classify haematological diseases with an accuracy up to 96%. We have collected, cleaned, preprocessed, and labelled an amount of 4124 hemograms from the Hospital Clínico San Carlos (Madrid, Spain) to be fed to our algorithm. Furthermore, we have used contribution analysis techniques to interpret our results and compare them with existing scientific knowledge.

The current global economic and technological perspective, has led to a decline in birth rates and an increase in the longevity of the population, resulting in an inverted population pyramid that is expected to worsen in the coming years [27]. Developing and investing in new methods to assist specialists in their tasks will improve the human health and the quality of life for patients, while optimising the healthcare system by saving time, money, and workload.

2. Methods

2.1. Data collection

Our data have been collected from the database of the Hospital Clínico San Carlos (Madrid, Spain). This study complies with all regulations, informed consent and was conducted according to the established ethical guidelines. The experiment was approved by the CEIC Hospital Clínico San Carlos. We obtained an amount of 114.789 hemograms from the haematological department. These samples were cleaned in order to standardise and anonymize the information. For this purpose, we removed the history number of the patient and the date they were tested. We conserved the birth date and the sex because of the relevance of this information for the diagnosis.

In order to assemble our dataset, we developed a Web Application, specially designed for mobile devices, to provide a labelling tool for the specialists. This tool enables doctors to label samples based on the information contained in the hemograms, peripheral blood smears and the genetic study performed (if applicable), which will be subsequently used in our supervised learning methods. We focused on 5 different categories (thalassemias, structural hemoglobinopathies, iron deficiency anaemia chronic disease anaemia and control healthy patients), which could be grouped according to their origin (healthy patients, congenital anaemias and acquired anaemias). Additionally, some subtypes of the disease have also been labelled regarding different types of thalassemias and structural hemoglobinopathies.

The reasons behind this specific selection of thalassemias and structural haemoglobinopathies are mainly the high prevalence of these pathologies in the territory of origin of the dataset (Spain), their observability in the hemogram and the common misclassification in their less severe form. In this way, the group of structural haemoglobinopathies is specially hard to identify with the information contained in a hemogram because its not linked to a specific value contained in it, but due to its high prevalence of this pathology in our territory (Spain), the reflection of this disease in the complete blood count (and its low variability over the years of

evolution) and its poor diagnosis (especially in the less severe forms) make them an interesting candidate to introduce into our dataset.

2.2. Data processing

We applied a cleaning process to remove outliers and corrupted samples and encode the information in a suitable format for artificial intelligence algorithms. Firstly, we transformed categorical features using one-hot encoding (e.g. using a binary field to represent Male as 0 and Female as 1). Then, we kept only the labelled samples separated by class and applied outlier filtering by columns based on z-scores, removing those that were at least 6 times away from the standard deviation. This decision was made by the doctors' team in order to detect cases where data was artifactualized but were not patients with real but extreme values.

We also calculated a correlation matrix using the Pearson correlation coefficient to identify correlated features. However, even with strong correlations (>0.8), we decided to keep the features because our hypothesis and preliminary studies suggested that the combination of some of those variables contributed to the final classification.

Finally, we generated a dataset where each sample is composed of 20 different features including 18 blood-related values and the sex and age of the patient (See *Table of features that were collected for the experiments* in the Supplementary material for more detail). The dataset comprises 4061 samples in total, labelled into 5 different categories and 7 subtypes of thalassemias and structural haemoglobinopathies.

2.3. Traditional algorithms: Support Vector Machines, Random Forests and boosting decision trees

In addition to neural networks, traditional artificial intelligence models were also deployed to serve as a comparison baseline. All our experiments were performed using Stratified K-Fold Validation with $K = 5$, where the train/test sets were split for each training process while maintaining class distribution, and a selection of hyperparameters based on grid search on the training set. Then, the models were trained with the best hyperparameter selection and evaluated on the test dataset. Due to the imbalance between classes in the dataset, we have also applied a weighted penalty based on the number of samples of each class in relation to the total number of samples. Three different methods were implemented: a) Support Vector Machine, b) Random Forest, c) Boosting Decision Trees.

In the first case, we have developed a Support Vector Machine experiment using the Support Vector Classification (SVC) implementation from the Scikit-Learn [28] library, which is based on LibSVM [29]. The grid search for the adjustment of our model has been done to tune three hyperparameters: a) The parameter C, which acts as a regularizer penalising a misclassified sample, b) The number of iterations needed for the method to get trained without overfitting the training data, and c) the kernel used to create the hyperplane, and d) the particular parameters of the selected kernel (i.e. gamma in the Radial Basis Function or the polynomial kernel). At first, a pre-screening for different kernels was done with default parameters, followed by a specific grid search for the hyperparameters as mentioned before.

Our second approach employs the Random Forest algorithm [30]. In this case, the grid search of hyperparameters to adjust the model has been done for three of them: a) The number of estimators (trees), b) the minimum number of samples required to be at a leaf node, c) the minimum number of samples required to split an internal node.

Finally, we have deployed boosting decision trees based on the XGBoost implementation for Python [31]. In this case, the list of hyperparameters we needed to tune was: a) the learning rate, b) the maximum depth of the trees, c) the number of estimators (trees), d) the minimum required to create a new node in the tree, e) subsample and f) column sample by tree, to determine the random subsample when a tree is created and at every new level, and g) the gamma parameters, which act as a regularizer across different trees to add nodes based on their contribution to the trees.

2.4. Neural networks - specific knowledge

As we have seen in the background section, there are few studies that are also limited to traditional machine learning models. Furthermore, the scope of the experiment is usually restricted to binary classification. We have designed a neural network architecture for both binary and multiclass classification, applying specific clinical knowledge and methodologies in three different areas:

1) *Architecture - Model input*: The data contained in the hemograms is fed to the network following a similar approach to that used by doctors to extract information from them: A) First, the information regarding the three different groups of data (red cells, white cells and platelets) are fed and passed through the input layer and distributed into different fully connected layers with ReLU [32] activations using lambda layers. B) The resulting activations are then concatenated and provided to the next block of fully connected layers, both of which also use ReLU activations. C) the full vector of features, including the sex and the age, is added at the end of the network as a residual connection. This approach is partially inspired by the design of the Cox-PASNet [33], where both clinical and genomic data is combined throughout the architecture. A total of 10 layers and 51 inputs are defined from the beginning to the end of the network. The interaction between them and the features is escalated and progressive as long as the information flows along it. A full image and description of the architecture can be seen in the *BloodNet Architecture* section in the Supplementary Material.
2) *Feature engineering:* We have created variables that specify if their related feature is within the ranges considered normal, or if it is high or low. If a sample has a value higher or lower than the range, it is indicated to the network through a value set in a new column. For example, in the case of haemoglobin .

3) *Contribution analysis*: Along with the performance metrics of the network (i.e. accuracy, precision and recall), we have used the Shapley Additive explanations (SHAP) framework [34] to weight the relevance of each feature by its contribution to the final decision. The idea behind this process is not only to understand why the network is making such a decision, but also to use this information in a future dashboard in order to also explain to doctors why this decision has been made for every new case (Fig. 2)

The proposed architecture also implements regularisers L2 and dropout that have been tuned along with the rest of hyperparameters to avoid overfitting. This prevents our model from memorising the training data, which will lead us to non-generalisable results. Additionally, callbacks for early stopping, which stop the training after a specific number of epochs when there are no improvements in the test set, and learning rate reduction on plateau have been employed in order to reduce the training phase dynamically without sacrificing our results. The list of hyperparameters to tune includes: a) hidden units ratio, b) batch size, c) learning rate and d) L2 rate.

2.5. Neural networks - general knowledge

State-of-the-art classification methods for tabular data using neural networks include approaches using attention [35] to improve their results. In this way, we wanted to compare our neural network with specific knowledge design against already tested neural network architectures. For this purpose, we have adapted the TabNet architecture [36], which uses sequential attention, to compare against our neural network (BloodNet) and the other traditional applied methods.

2.6. Experiment definition

Three different experiments have been defined in order to compare the performance of our different models: a) a binary classification between healthy patients and diseased, b) a multiclass classification including the 3 big groups contained in the dataset (healthy patients, congenital anaemias and acquired anaemias), and c) a multiclass classification including all the subgroups (healthy patients, Thalassaemias, Structural haemoglobinopathies, Iron deficiency anaemia, Chronic disease anaemia).

Stage	Name	Group
1	Binary classification	Healthy and diseased
2	Multiclass classification (3 groups)	Healthy, congenital anaemias and acquired anaemias.
3	Multiclass classification (5 groups)	Healthy, thalassaemias, structural haemoglobinopathies, iron deficiency anaemia and chronic disease anaemia

As mentioned before, structural haemoglobinopathies are very difficult to identify with the information contained in a regular haemogram. The purpose of adding them to the group of diseases was to also study their effect on the rest of the classification and how the performance was affected. For this purpose, we replicated the same experiments but removed them from the dataset.

For each experiment, accuracy, precision and recall were collected. To obtain such values, we used a 5-Fold stratified cross-

Table 1
Results for the binary classification experiment. It contains the value of the accuracy, precision and recall by class and algorithm expressed as a percentage. In terms of accuracy, the best performing algorithm, in both cases, is the neural network, followed closely by the random forest (RF). The same applies to precision and recall. SVM stands for Support Vector Machine.

Accuracy				
Algorithm	With Structural Haemog.		Without Structural Haemog.	
SVM	86,1		92,8	
RF	88,9		95,9	
XGboost	86,6		94,1	
BloodNet	89,1		96,4	
TabNet	85,7		92,9	
Precision				
	With Structural Haemog.		Without Structural Haemog.	
	Healthy	Diseased	Healthy	Diseased
SVM	63	96	91	94
RF	73	92	93	96
XGboost	70	92	93	94
BloodNet	71	97	93	98
TabNet	69	93	90	94
Recall				
	With Structural Haemog.		Without Structural Haemog.	
Algorithm	Healthy	Diseased	Healthy	Diseased
SVM	89	83	90	95
RF	78	90	93	96
XGboost	77	88	88	96
BloodNet	93	86	96	96
TabNet	82	87	89	95

validation, followed by a performance review using the test set. To deal with the randomness of the models, we took the mean of a total of 10 runs per experiment. Furthermore, and due to the imbalance of the different classes of the dataset, weights per class were applied during the training phase to penalise the most represented classes and avoid bias. The specific hyperparameters used for each of the experiments are provided in the *List of hyperparameters for the proposed models* in the Supplementary Material. Specific examples of ROC, AUROC, PRC, AUPRC and confusion matrices are also available in the Supplementary Material under the *Metrics* section due to the large amount of values.

3. Results

3.1. Stage 1

The first experiment (see Table 1) consists of a binary classification where all the labelled data was split in two different groups: 1) healthy patients, which were already a class in our dataset and 2) the rest of the samples were merged in a general class to group the different haematological diseases.

3.2. Stage 2

In this experiment (see Table 2) we have grouped our classes using the hierarchically superior family of the different labelled diseases. We have 1) a group that contains the healthy patients, 2) a group containing diseases from the family of Congenital Anaemias and 3) a group for the family of Acquired Anaemias.

3.3. Stage 3

For our last experiment (see Table 3), we have split our dataset in 5 different groups, each of them containing a specific class labelled in our dataset: 1) healthy patients, 2) Thalassaemias, 3) Structural Haemoglobinopathies, 4) Iron deficiency anaemia, 5) Chronic disease anaemia.

3.4. Contribution analysis

We have performed a contribution analysis using SHAP. For this purpose we have run two different tests in order to explain the behaviour of our neural network. As a reference, we have used the 3-class experiment model for the study.

Table 2
Results for the 3-class classification experiment. It contains the value of the accuracy, precision and recall by class and algorithm expressed as a percentage. In terms of accuracy, the best performing algorithm, in both cases, is the neural network, followed closely by the random forest. The best precision and recall, in this case, are distributed between the neural network, the XGboost and the random forest (RF). SVM stands for Support Vector Machine.

Accuracy						
Algorithm	With Structural Haemog.			Without Structural Haemog.		
SVM	74,6			87,3		
RF	74,0			89,6		
XGboost	74,1			87,9		
BloodNet	78,0			90,1		
TabNet	71,9			86,3		
Precision						
	With Structural Haemog.			Without Structural Haemog.		
	Healthy	Congenital	Acquired	Healthy	Congenital	Acquired
SVM	66	86	74	88	80	88
RF	77	77	68	93	87	89
XGboost	72	79	76	88	90	87
BloodNet	75	88	74	93	86	88
TabNet	68	78	69	91	82	85
Recall						
	With Structural Haemog.			Without Structural Haemog.		
	Healthy	Congenital	Acquired	Healthy	Congenital	Acquired
SVM	87	59	87	89	86	83
RF	85	62	81	94	87	88
XGboost	90	65	80	91	81	90
BloodNet	88	66	88	90	88	91
TabNet	80	61	80	89	82	86

Table 3
Results for the 5-class classification experiment. It contains the value of the accuracy, precision and recall by class and algorithm expressed as a percentage. The best performing algorithm is the neural network followed closely by the random forest (RF) and the XGBoost. The best precision and recall, in the case of the experiment including haemoglobinopathies, is for the neural network, while in the case without them, it is very irregular and distributed between the different algorithms. SVM stands for Support Vector Machine.

Accuracy									
Algorithm	With Structural Haemog.					Without Structural Haemog.			
SVM	66,9					82,4			
RF	67,6					87,1			
XGboost	68,0					84,3			
BloodNet	71,8					88,6			
TabNet	65,0					82,4			
Precision									
	With Structural Haemog.					Without Structural Haemog.			
	Healthy	Thalassemias	Structural H.	Iron Def. A.	Chronic D. A.	Healthy	Thalassemias	Iron Def. A.	Chronic D. A.
SVM	67	74	56	66	67	90	89	69	72
RF	69	74	53	64	70	92	87	75	87
XGboost	72	74	53	68	67	90	85	79	81
BloodNet	76	76	69	73	59	87	87	81	85
TabNet	62	69	34	60	69	88	85	71	79
Recall									
	With Structural Haemog.					Without Structural Haemog.			
	Healthy	Thalassemias	Structural H.	Iron Def. A.	Chronic D. A.	Healthy	Thalassemias	Iron Def. A.	Chronic D. A.
SVM	81	81	25	73	85	90	79	71	83
RF	86	79	30	71	77	95	84	71	93
XGboost	85	81	34	60	81	92	84	69	90
BloodNet	93	89	34	75	72	88	86	69	93
TabNet	83	77	22	56	62	93	82	67	79

4. Discussion and conclusions

The use of artificial intelligence for data analysis has proven to be a very effective method for discovering new insights. The clinical field should not fall behind this trend. Clinical data has its own characteristics and limitations: on one hand, intrinsic privacy makes access, collection and use a challenge in itself. On the other hand, the lack of standardisation, the diversity in methodologies, and the state of digitalisation of hospitals make it more complicated to progress.

Regarding the first point, Federated Learning seems to be a feasible solution. Even with its own limitations, this technology is bringing up new ideas to use artificial intelligence without putting privacy at risk, while saving time and administrative costs. For the second point, it appears that the awareness of the potential improvements of using advanced data analysis techniques, such as Deep Learning, is leading to an interest in improving every step of the data protocol. Therefore, the exploration and proposals of new methodologies that help in this process are necessary.

While the use of neural networks is becoming more common regarding images or sequential data, the results are not as good when it comes to tabular data. The complexity of the algorithms, the difficulty of tuning, and the lack of large amounts of data are the main drawbacks. On the other hand, the flexibility of neural networks offers us the opportunity to use the specific knowledge (as we describe in the methods section) of the use case in its design to improve the results.

Regarding our results, in the binary classification experiment (stage 1), using our neural network, we are obtaining 89.1% and 96.4% in accuracy when including and removing hemoglobinopathies respectively. We can observe that the differences in accuracy between our best algorithm and the worst (including traditional methods) is 3.4%, being the mean 2.2%, when including structural hemoglobinopathies, and 3.6% and 2,4% when removing them.

In our 3-class classification experiment (stage 2), the best performer is also the neural network obtaining 78% and 90.1% in accuracies with a maximum difference of 6.1% and a mean difference of 4.3% in the case of including the structural hemoglobinopathies, and 3.8% and 2.3% in the case of removing them (see Fig. 1).

Finally, for the 5-class classification experiment (stage 3) the accuracy obtained is 71.8%, with a maximum difference of 6.8% and a mean difference between algorithms of 4.9%. We can observe that both precision and recall are also higher and stable in the experiment using our neural network.

Therefore, our results demonstrate that neural networks are performing better than the others algorithms. This could be due to the use of techniques to include specific knowledge of the problem, as the comparison with the TabNet (without specific knowledge) suggests, but further research needs to be done. We expected to obtain better results with the TabNet architecture, and neural network approaches in general, but the lack of data may be a significant limitation which could explain this behaviour. Regarding specific knowledge neural networks architectures, another of their limitations is that they should probably be adapted in terms of data variations (i.e. data quantity, data quality or features and class distribution) which requires a complete training and validation workflow. Furthermore, our dataset was specific to a certain geographical area, which may limit the applicability of the trained model but will allow more experiments using transfer learning and fine-tuning techniques.

In our opinion, including and taking advantage of this knowledge is far more accessible in neural network developments. However,

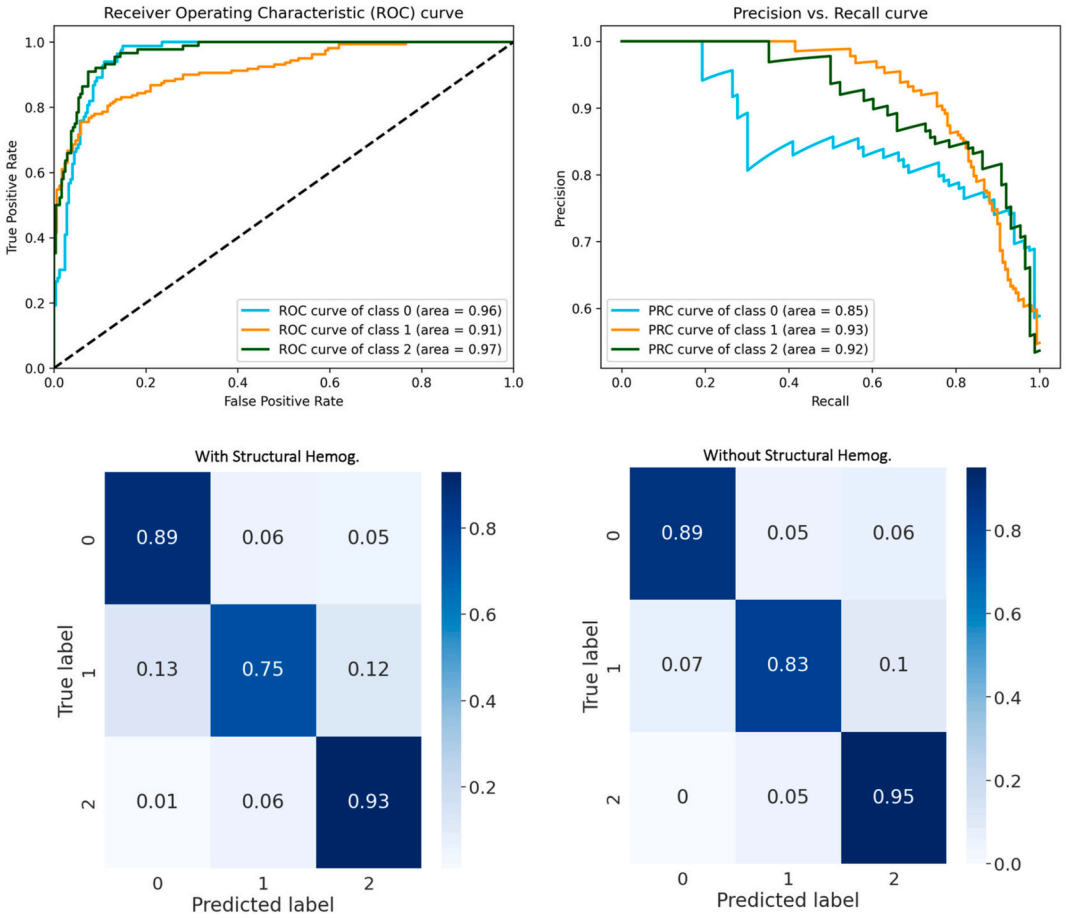

Fig. 1. On the top, the ROC curve and the PRC curve along with the AUROC and AUPRC for the stage 2 experiment, are shown. Class 0 represents healthy patients; Class 1 represents Congenital Anaemias and Class 2 represents Acquired Anaemias. On the bottom, the confusion matrix for the same experiment (see also Table 2) is displayed. Additional information regarding ROC, AUROC, PRC, AUPRC and confusion matrices are also available in the Supplementary Material under the *Metrics* section.

the time spent in the design, development, training and tuning of these algorithms is also more demanding and consumes more resources than the others. In this sense, a preliminary study of the use case should be done in order to determine the importance of the outcome in comparison with the complexity of the problem and the resources needed. In our opinion, there will be cases where this increase in resources would not be worthy from a project management point of view, so there will be alternatives for a better trade-off.

We have also seen how it is mandatory to organise a framework of collaboration and continuous communication from the beginning to the end of the pipeline. From data collection to the understandability of our model, the healthcare practitioners' knowledge has been essential. The use of different sources of data could help in avoiding bias regarding classification problems. In this way, Federated learning seems to be one of the best alternatives to share knowledge without sharing the data, which is one of the main limitations regarding the use of neural networks in healthcare problems.

These tools should be developed taking into account that they are going to be used to assist specialists in order to reduce their workload while speeding up diagnoses with very high accuracy (up to 96% in our work), but never with the aim of replacing them. Therefore, further research in visualisation and bench-to-bed transition for AI models is becoming mandatory to take advantage of the results.

Finally, the explainability of our models is as important as the results themselves. This is even more important in biomedicine because, in order to assist correctly, specialists should know why the machine is providing that decision [37,38], so 1) we can confirm with the literature that decision being taken has a scientific background and 2) that the doctor can take that information into account. In this way, the use of tools like SHAP eases the explainability of our models. In our case, we can see in Fig. 2 that, for the specific case of thalassemias, a low value of MCH and a high value of HEMA are the main reasons for classifying a sample as a thalassemia [39,40].

It is worth noting that artificial Intelligence models are highly sensitive to the data they are trained with and how the training phase is conducted. Therefore, explainability of the models should be considered an essential part of any validation pipeline regarding artificial intelligence. This is even more critical in a clinical environment. Moreover, in order to increase the confidence of the clinicians and the robustness of our models, validation steps must be implemented throughout the complete production workflow (i.e.

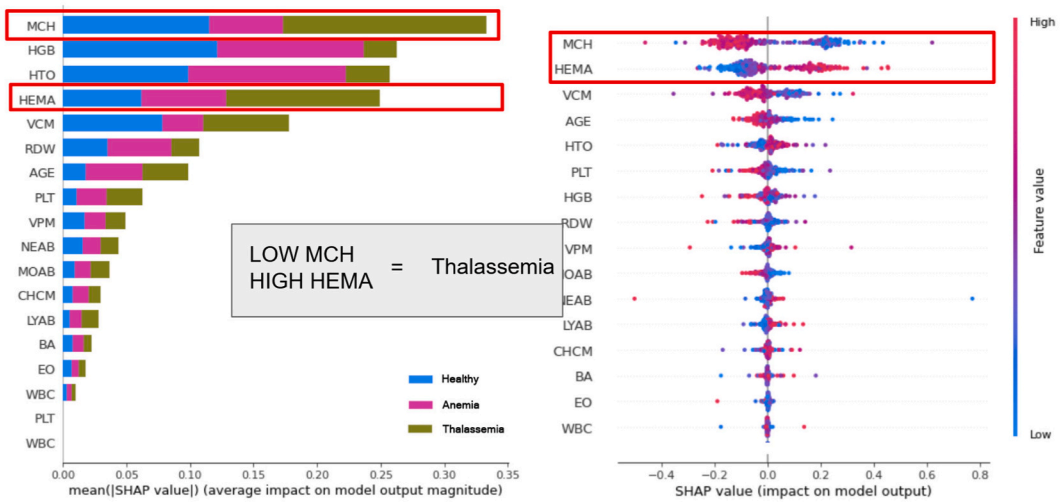

Fig. 2. On the left, a deep explainer plot from the SHAP framework is shown. In this plot, we can see the most important features used to classify our samples. We can see that, for Thalassaemias, the most important features are the MCH (mean corpuscular haemoglobin) and the HEMA (hematies) values, which belong in both cases to the red blood cells group. In the right, a summary plot for the specific class of Thalassaemias. Here we can see how the values of the feature have affected the classification for this specific class. We can see that low values of MCH, and high values of HEMA, are very important characteristics used to classify a sample as a thalassaemia. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

data drifting methods, model performance downgrades, data changes or human-in-the-loop strategies).

Artificial Intelligence allows us to gain new insights from data, but the collaboration between different fields is becoming mandatory in order to take advantage of the potential of these tools. Specifically with clinical data, the interdisciplinary collaboration between engineers and healthcare practitioners could lead us to the development of improved algorithms to solve problems using specific knowledge, as we have shown. This process starts at the very beginning, from data collection and labelling, to making use of the predictions as systems that aid practitioners in real environments, so we can make a difference in the quality of life for people while improving the workload of the doctors, making our healthcare system more efficient.

Funding

This work has been partially supported by the European project (grant no. 676559) (European Union), the Spanish national project Plataforma de Recursos Biomoleculares y Bioinformáticos (ISCIII-PT13.0001.0012 and ISCIII-PT17.0009.0022) (Spain), the Fondo Europeo de Desarrollo Regional (UMA18-FEDERJA-156, UMA20-FEDERJA-059) (Andalucia, Spain), the Junta de Andalucía (P18-FR-3130) (Andalucia, Spain), the Instituto de Investigación Biomédica de Málaga IBIMA and the University of Málaga (Spain).

Author contribution statement

Sergio Diaz-del-Pino; Roberto Trelles-Martinez; Ataulfo Gonzalez; Nicolas Guil: Conceived and designed the experiments; Performed the experiments; Analysed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Data availability statement

The authors do not have permission to share data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Acknowledgements

We wish to thank the BITLAB Group, the Hospital Clinico San Carlos and the application testers for all their support and comments which have significantly contributed to improve this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e15940.

References

[1] Statista Search Department, Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 [Infographic]. https://www.statista.com/statistics/871513/worldwide-data-created/, 2022. (Accessed 22 February 2023) accessed.

[2] F.G. dos Santos, Claudio, D.S. Oliveira, L.A. Passos, R. Gonçalves Pires, D. Felipe Silva Santos, L. Pascotti Valem, T.P. Moreira, M. Cleison S. Santana, M. Roder, J. Paulo Papa, D. Colombo, Gait recognition based on deep learning: a survey, ACM Comput. Surv. 55 (2023), https://doi.org/10.1145/3490235. Article 34.

[3] F.M. Castro, M.J. Marín-Jiménez, N. Guil, N. Pérez de la Blanca, Automatic Learning of Gait Signatures for People Identification, in: I. Rojas, G. Joya, A. Catala (Eds.), Advances in Computational Intelligence, Springer International Publishing, Cham, 2017, pp. 257–270, https://doi.org/10.1007/978-3-319-59147-6_23.

[4] Q. Rao, J. Frtunikj, Deep Learning for Self-Driving Cars, in: Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems, Gothenburg Sweden, Association for Computing Machinery, New York, NY, USA, 2018, pp. 35–38, https://doi.org/10.1145/3194085.3194087.

[5] S.A. Fahad, A.E. Yahya, Inflectional review of deep learning on natural language processing, in: 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), Shah Alam, IEEE, New Jersey,N J, USA, pp. 1–4. https://doi.org/10.1109/ICSCEE.2018.8538416.

[6] I.I. Baskin, D. Winkler, I.V. Tetko, A renaissance of neural networks in drug discovery, Expet Opin. Drug Discov. 11 (2016) 785–795, https://doi.org/10.1080/17460441.2016.1201262.

[7] A. Eetemadi, I. Tagkopoulos, Genetic neural networks: an artificial neural network architecture for capturing gene expression relationships, Bioinformatics 35 (2019) 2226–2234, https://doi.org/10.1093/bioinformatics/bty945.

[8] S. Agrawal, J. Agrawal, Neural network techniques for cancer prediction: a survey, Procedia Comput. Sci. 60 (2015) 769–774, https://doi.org/10.1016/j.procs.2015.08.234.

[9] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (2019) 44–56, https://doi.org/10.1038/s41591-018-0300-7.

[10] T. Alsuliman, D. Humaidan, L. Sliman, Machine learning and artificial intelligence in the service of medicine: necessity or potentiality? Curr. Res. Transl. Med. 68 (2020) 245–251, https://doi.org/10.1016/j.retram.2020.01.002.

[11] M. Mallappallil, J. Sabu, A. Gruessner, M. Salifu, A review of big data and medical research, SAGE Open Med. 8 (2020), 205031212093483, https://doi.org/10.1177/2050312120934839.

[12] N. Radakovich, M. Nagy, A. Nazha, Artificial intelligence in hematology: current challenges and opportunities, Curr. Hematol. Malig. Rep. 15 (2020) 203–210, https://doi.org/10.1007/s11899-020-00575-4.

[13] L. Kaestner, Artificial intelligence meets haematology, Transfus. Apher. Sci. 59 (2020), 102986, https://doi.org/10.1016/j.transci.2020.102986.

[14] L. Kaestner, P. Bianchi, Trends in the development of diagnostic tools for red blood cell-related diseases and anemias, Front. Physiol. 11 (2020), https://doi.org/10.3389/fphys.2020.00387.

[15] K.Y. Leung, C.P. Lee, M.H.Y. Tang, E.T. Lau, L.K.L. Ng, Y.P. Lee, H.Y. Chan, E.S.K. Ma, V. Chan, Cost-effectiveness of prenatal screening for thalassaemia in Hong Kong, Prenat. Diagn. 24 (2004) 899–907, https://doi.org/10.1002/pd.1035.

[16] J.V. Caballero, Cromatografía Líquida Desnaturalizante De Alto Rendimiento (Dhplc): Uso Y Utilidad En El Genotipado Masivo [Denaturing High Performance Liquid Chromatography (Dhplc): Use and Utility in Massive Genotyping: "Executive Abstract"], Agencia de Evaluación de Tecnologías Sanitarias de Andalucía, Sevilla, 2008.

[17] F. Yousefian, T. Banirostam, A. Azarkeivan, Prediction thalassemia based on artificial intelligence techniques: a survey, Int. J. Adv. Res. Comput. Commun. Eng. 6 (2017) 1–3, https://doi.org/10.17148/IJARCCE.2017.6847.

[18] A.S. AlAgha, H. Faris, B.H. Hammo, A.M. Al-Zoubi, Identifying β-thalassemia carriers using a data mining approach: the case of the Gaza Strip, Palestine, Artif. Intell. Med. 88 (2018) 70–83, https://doi.org/10.1016/j.artmed.2018.04.009.

[19] S.R. Amendolia, G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala, G.M. Mura, A comparative study of k-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening, Chemometr. Intell. Lab. Syst. 69 (2003) 13–20, https://doi.org/10.1016/S0169-7439(03)00094-7.

[20] F. Akter, M.A. Hossin, G.M. Daiyan, M.M. Hossain, Classification of haematological data using data mining technique to predict diseases, J. Comput. Commun. 6 (2018) 76–83, https://doi.org/10.4236/jcc.2018.64007.

[21] W.S. Noble, What is a support vector machine? Nat. Biotechnol. 24 (2006) 1565–1567, https://doi.org/10.1038/nbt1206-1565.

[22] A. Cutler, D.R. Cutler, J.R. Stevens, Random Forests, in: C. Zhang, Y. Ma (Eds.), Ensemble Machine Learning, Springer US, Boston, MA, 2012, pp. 157–175, https://doi.org/10.1007/978-1-4419-9326-7_5.

[23] H. Taud, J.F. Mas, Multilayer perceptron, in: M.T. Camacho Olmedo, M. Paegelow, J.-F. Mas, F. Escobar (Eds.), Geomatic Approaches for Modelling Land Change Scenarios, Springer International Publishing, Cham, 2018, pp. 451–455, https://doi.org/10.1007/978-3-319-60801-3_27.

[24] S. Vijayarani, S. Sudha, An efficient clustering algorithm for predicting diseases from hemogram blood test samples, Indian J. Sci. Technol. 8 (2015) 1.

[25] G. Gunčar, M. Kukar, M. Notar, M. Brvar, P. Černelč, M. Notar, M. Notar, An application of machine learning to haematological diagnosis, Sci. Rep. 8 (2018) 411, https://doi.org/10.1038/s41598-017-18564-8.

[26] Z. Zhang, M.W. Beck, D.A. Winkler, B. Huang, W. Sibanda, H. Goyal, Opening the black box of neural networks: methods for interpreting neural network models in clinical applications, Ann. Transl. Med. 6 (2018) 216, https://doi.org/10.21037/atm.2018.05.32.

[27] M. Cristea, G.G. Noja, P. Stefea, A.L. Sala, The impact of population aging and public health support on EU labor markets, Int. J. Environ. Res. Publ. Health 17 (2020) 1439, https://doi.org/10.3390/ijerph17041439.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[29] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27.

[30] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition 1, Canada, Montreal, 1995, pp. 278–282.

[31] T. Chen, C. Guestrin, XGBoost, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA, ACM, New York, NY, USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[32] A.F. Agarap, Deep Learning Using Rectified Linear Units, 2018. Arxiv Preprint Arxiv.

[33] J. Hao, Y. Kim, T. Mallavarapu, J.H. Oh, M. Kang, Cox-pasnet: pathway-based sparse deep neural network for survival analysis, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, IEEE, New Jersey,N J, USA, pp. 381–386.

[34] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, NeurIPS, 2017, pp. 30–40.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[36] S.Ö. Arik, T. Pfister, Tabnet: Attentive Interpretable Tabular Learning, 2019 arXiv, Arxiv Preprint Arxiv.

[37] A. Holzinger, The Next Frontier: AI We Can Really Trust. Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2021. Communications in Computer and Information Science, vol 1524. Springer, Cham. https://doi.org/10.1007/978-3-030-93736-2_33.

[38] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J.D. Ser, W. Samek, I. Jurisica, N. Díaz-Rodríguez, Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Inf. Fusion 79 (2022) 263–278, https://doi.org/10.1016/j.inffus.2021.10.007.
[39] B. Arrizabalaga Amuchástegui, F.A. González Fernández, Á.F. Remacha Sevilla, Eritropatología, Ambos Marketing Services, Barcelona, 2017.
[40] V. Brancaleoni, E. Di Pierro, I. Motta, M.D. Cappellini, Laboratory diagnosis of thalassemia, Int. J. Lab. Hematol. 38 (Suppl 1) (2016) 32–40, https://doi.org/10.1111/ijlh.12527.