IJSES
Volume 9, Issue 10, pp. 16-23, 2025.
International Journal of Scientific Engineering and Science
ISSN (Online): 2456-7361

Classification of Abo Blood Groups Via Spectroscopic Analysis and Machine Learning: Optimization of a Predictive Model for Antigen A

Fatima Ezzahra El Kamouny[1], Younes Wadiai[2], Ayoub El Idrissi[3], Abdellah Madani[4], Khiat Amina[5]

[1]Laboratory LAROSERI, Department of Computer Science, University Chouaib Doukkali, Faculty of Sciences, B.P. 20,24000 El Jadida, Morocco.

[2]Laboratory of Innovative Systems Engineering, National School of Applied Sciences of Tetouan, Abdelmalek Essaâdi University, Tetouan, Morocco.

[3]National School of Applied Sciences, University Chouaib Doukkali, B.P. 20,24000 El Jadida, Morocco.

[4]Laboratory LAROSERI, Department of Computer Science, University Chouaib Doukkali, Faculty of Sciences, B.P. 20,24000 El Jadida, Morocco.

[5]Laboratory LAROSERI, Department of Computer Science, University Chouaib Doukkali, Faculty of Sciences, B.P. 20,24000 El Jadida, Morocco.

E-mail: [1]f.elkamouny@gmail.com, [2]y.wadiai@uae.ac.ma, [3]e-elidrissi.a@ucd.ma, [4]madani.a@ucd.ac.ma, [5]Khiat.amina@ucd.ac.ma

Abstract— *Accurate blood group determination is crucial for transfusion safety. Traditional methods mainly use immunohematological tests, which can be costly, invasive, and reliant on human interpretation. This study introduces an alternative method using near-infrared (NIR) spectroscopy analysis of red blood cells combined with machine learning algorithms for ABO blood group classification. An experimental protocol was performed on 211 blood samples, with spectra collected from 1000 to 2600 nm. Three models were evaluated to predict the presence of antigen A: Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN). The results show that the ANN model achieved the highest accuracy (98.9%), followed by SVM (98.7%) and RF (96%). These results demonstrate the potential of spectroscopic and intelligent approaches for rapid, noninvasive, and portable blood typing, especially in out-of-hospital and emergency settings.*

Keywords— *ABO blood typing; Antigen A; NIR spectroscopy; Machine learning; SVM; Random Forest; Neural Network; Classification; Transfusion Medicine.*

I. INTRODUCTION

The safe implementation of blood transfusion primarily relies on accurate blood group identification. The ABO system, which was first identified by Karl Landsteiner in 1901, is still the basis of blood type matching that is used today. This blood group system, present on the surface of red cells, is responsible for the existence or absence of A and B antigens and the corresponding A and B natural antibodies in plasma.Acute haemolytic reactions and other serious complications, such as transfusion shock, disseminated intravascular coagulation and acute renal failure, or even patient death, could result if transfusions with incompatible blood are given [1-2]. Blood typing, therefore, is an essential part of transfusion medicine, and is usually performed prior to a transfusion, or any significant surgical procedure, as well as before organ transplantation. In the past, blood groups were determined using standard immunohematological methods utilizing antibody-antigen reactions. The most commonly used in practice are manual agglutination tests, which include the addition of the patient's red blood cells to anti A and anti B antibodies which are present in sera. Entails: Looking for agglutination which is That helps indicate the blood type. To be announced, the group. There are 2 basic types: slide agglutination (Beth- Vincent method) and tube agglutination; the latter is more sensitive. and able to find weak harmonics as A 2 or A χ [3]. These tests are simple to perform but extremely subjective (visually), non-reproducible.

Quality control was enhanced in the 1990s by adding semi-automated methods such as gel columns. These facilitate the selective migration of erythrocytes through gel containing antibodies, by centrifugation. The location of the cells in the column following the reaction indicates whether or not agglutination has occurred. Concurrently, low ionic strength solutions or the indirect Coombs test were introduced for better recognition of irregular antibodies and for Rh typing [4-5]. More recently, automatic dispensers have been used in great hospital centers. Among them, the use of flow cytometry, based on fluorescent antibodies detecting surface proteins, allows a rapid and refined multiparametric analysis. Some other systems include ECL in which antigen- antibody reactions are detected by an amount of light in proportion to the bonding generated. These procedures are highly reliable but depend on costly instrumentation, trained technical service and a controlled environment [6-7-8]. However, these conventional approaches seemingly have some primary limitations. They are invasive, as venous sampling is required, which is not always feasible in

The emergency or rural environment. In addition, costly and sensitive biological reagents are used, creating supply chain and

stability issues. Furthermore, manual techniques rely much on human interpretation, which brings the risk of subjective mistakes. Finally, they need laboratory setup equipment and skilled staff, thus making it impossible to be used in resource-poor locations or in case of emergency [12-11-10]. In light of this, biomedical research is currently investigating technological substitutes that can overcome these limitations. Artificial intelligence (AI) and spectroscopy are two prominent innovation avenues. The vibrational signature of biomolecules found on the outside or inside of red blood cells can be analyzed using Fourier-transform infrared (FTIR) and Raman spectroscopy. Each sample receives a distinct spectrum from these non-destructive, reagent-free methods that represent its overall biochemical makeup [13–14]. Nevertheless, processing these intricate and multifaceted spectra calls for strong instruments. This spectral data can be effectively used with the help of artificial intelligence. Support Vector Machines (SVM), Random Forests, and Convolutional Neural Networks (CNN) are examples of machine learning algorithms that have shown they can recognize distinctive patterns in near-infrared (NIR) spectra, frequently with an accuracy of over 95% [15-16-17]. The creation of automated, replicable, and expandable systems is made possible by these models. Without the need for invasive sampling or trained staff, recent prototypes that combine FTIR spectroscopy and cellphones have been successfully tested for ABO typing, yielding findings in less than 10 seconds [18]. The current study intends to investigate the viability of a blood type system based on spectroscopic examination of red blood cells in conjunction with machine learning models in this developing technological setting. More precisely, it seeks to create a model that can identify antigen A from infrared spectra and assess its performance by contrasting it with reference methods in terms of accuracy, sensitivity, and specificity. The ultimate goal of this research is to suggest a non-invasive, portable, quick, and affordable solution that can be used outside of hospitals, in emergency situations, or in medically disadvantaged areas.

II. Machine Learning

A subfield of artificial intelligence called machine learning [19] uses statistical techniques to let computer systems learn from data. It is mostly predicated on three essential steps:

- Using a large dataset for training,
- Being able to extrapolate to new circumstances,
- Making forecasts based on fresh facts.

Three supervised learning techniques SVM, RF, and ANN were applied in this study to identify the presence of antigen A in a blood drop. The efficiency of SVMs [20] and ANNs [21] in the biomedical domain is highlighted in a large number of research papers. Notably, these models are utilized for computer-aided diagnosis of a number of diseases, including sleep apnea [24], diabetic retinopathy [23], and cardiac abnormalities [22].The three models that were tested SVM, Random Forest, and ANN as well as the metrics that are frequently used to evaluate their performance are summarized in the sections that follow. To enable an unbiased evaluation of the predicted accuracy of each algorithm, they were all trained on the same dataset. Because the input data (spectral measurements) are linked to known output labels (antigen A presence or absence), this work is classified under supervised learning. Thus, to guarantee the efficacy of the created models, a varied and pertinent dataset is essential. There are Random Forests, and Convolutional Neural Networks (

CNN) are examples of machine learning algorithms that have shown they can recognize distinctive patterns in near-infrared (NIR) spectra, frequently with an accuracy of over 95% [15-16-17]. The creation of automated, replicable, and expandable systems is made possible by these models. Without the need for invasive sampling or trained staff, recent prototypes that combine FTIR spectroscopy and cellphones have been successfully tested for ABO typing, yielding findings in less than 10 seconds [18]. The current study intends to investigate the viability of a blood type system based on spectroscopic examination of red blood cells in conjunction with machine learning models in this developing technological setting. More precisely, it seeks to create a model that can identify antigen A from infrared spectra and assess its performance by contrasting it with reference methods in terms of accuracy, sensitivity, and specificity. The ultimate goal of this research is to suggest a non-invasive, portable, quick, and affordable solution that can be used outside of hospitals, in emergency situations, or in medically disadvantaged areas. numerous parameters in machine learning algorithms; some of them are set prior to training, while others are learned during the training process. them are known as hyperparameters. The performance of the model is greatly impacted by the hyperparameter optimization.

2.1 Support Vector Machine

One supervised learning method that is well known for its resilience and capacity to manage regression and classification issues is the Support Vector Machine (SVM) [26]. By using kernel functions, this technique can take a linear or nonlinear approach, enabling the modeling of intricate interactions between data [27]. Finding the ideal hyperplane that maximizes the margin between points from distinct classes is the basic idea behind Support Vector Machines (SVM), which enhances the model's capacity for generalization. By serving as decision boundaries, these hyperplanes make it easier to classify fresh observations accurately. [28]. SVM can be applied to regression issues, where it aims to fit an approximation function by minimizing errors, even if its primary use is in classification [29]. The parameter C regulates the model's regularization, striking a balance between the model's intricacy and its capacity to withstand classification errors. While a high value of C limits errors at the risk of overfitting, a low value supports a wider margin by permitting more errors [30]. The Radial Basis Function (RBF) kernel is commonly used to handle nonlinear data. utilized. With the use of this kernel, the data can be projected onto a higher-dimensional space, enabling a linear separation. Its purpose is outlined by:

$$k(x, x') = \exp(-\gamma(x - x')^2) \quad (1)$$

where γ gamma (gamma) is a hyperparameter that regulates the local influence of a training vector, and.$|x - x'|$. is the squared Euclidean distance between two locations x and x'. While very low values of γ gamma may underestimate the complexity of the data, high values might cause overfitting by making the

decision boundary too sensitive to individual data points [31]. For the classifier to perform as well as feasible, it is essential to simultaneously adjust the parameters C and γ gamma in order to achieve an ideal trade-off between bias and variance [32].

2.2 Random Forest

The Random Forest (RF) algorithm is a machine learning method that combines many decision trees in a supervised way. Introduced by Leo Breiman in 2001 [33], it merges two key techniques: bagging and the random selection of variable subsets for each split. This combined approach reduces overfitting, even with noisy or highly correlated data, while also lowering the model's variance and improving its ability to generalize [34]. The core concept of RF is to create a collection of independent decision trees, each built from a bootstrapped echantillon of the learning game. To find the best separation, a sub- ensemble aléatoire of variables is chosen at each noeud during the development of each tree. This process introduced structural diversity among the trees, which significantly increased the overall robustness. After the fortress is constructed, the final prediction is made by either a majority vote in the case of a classification or a mean vote in the case of a regression. Furthermore, because of the nature of the bootstrap, approximately one layer of data—known as out-of-bag, or OOB—is not used in the creation of a given tree. Thus, this data can be used to assess the model's performance without the need for explicit cross-checking [35]. The RF offers a number of benefits, including the ability to handle large data sets, naturally handle complicated variable interactions, provide measures of attribute importance, and remain relatively insensitive to anomalous values. However, it may be less interpretable than a unique decision tree, and if no corrective measures are taken, its effectiveness may decline in cases of very unbalanced classes [36].

2.3 Artificial neural networks

Artificial neural networks, or ANNs, are a significant class of automatic learning systems. Inspired by the functioning of biological neural networks, these models are made up of interconnected treatment units called artificial neurons that enable modeling complex non-linear relationships between input and output variables [37]; artificial neural networks (ANNs) are data- oriented algorithms that learn from examples and attempt to extract statistically significant relationships without explicitly stating the underlying laws of the body [38]. Each neuron performs a linear combination of its inputs, to which a bias is added, followed by a non-linear activation function, introducing the ability to model complex decision boundaries. In our work, we have used a feedforward architecture, which is one of the most traditional topologies in supervised classification [38].This kind of network is organized into successive couches, as seen in Figures 1 and 2: A couche d'entrée, which receives spectral characteristics (such as onde lengths); one or more couches cached, which represent non-linear signaux combinations; and a couche de sortie, which, with the aid of the softmax activation function, produces a probability distribution across the classes. especially well-suited for multi-class classification tasks. Information travels in a single direction, from entry to exit, without going back. The network's output is compared to the expected output at each learning iteration (epoch); the error is then retropropagated inside the network using the backpropagation algorithm, which relies on gradient descent to adjust synaptic weights [39]. The often used loss function is entropie croisée, which is especially well-suited for multi-class classification tasks. Learning proceeds iteratively until it reaches a stopping criterion (a fixed number of epochs, a cost function that stagnates, or a target level of precision). Notwithstanding the computational costs of this stage, a well-enrolled network has a strong capacity for generalization on new data. It is important to note that there is no universal methodology for determining the optimal network design; instead, empirical definitions of the number of couches, neurons, activation functions, or learning rate are typically required. This approach involves trial and error, starting with a basic framework and adjusting the complexity based on observed performances [40]. Once a satisfactory model has been obtained, hyperparameter optimization techniques may be applied to further enhance it.

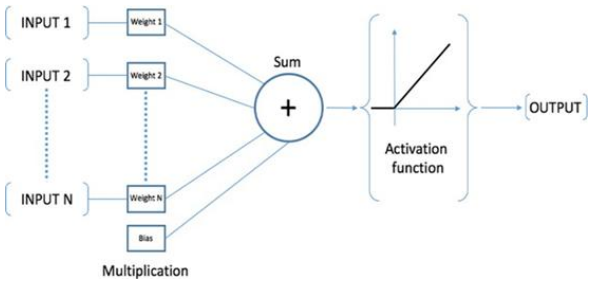

Figure 1: Working principle of an artificial neuron.

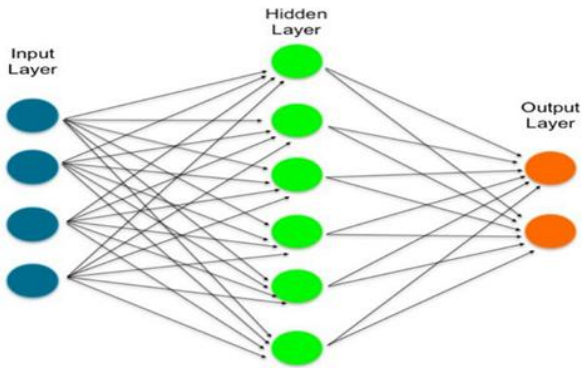

Figure 2: Chematic structure of feedforward propagation.

2.4 Metrics Used

A crucial step in any supervised learning process is the thorough evaluation of classification model performance, especially in binary classification tasks like anomaly detection. The goal of this investigation is to predict the presence or absence of antigen A based on spectroscopic data. Several metrics derived from the confusion matrix have been used to objectively measure the accuracy of predictions made by the tested algorithms (SVM, Random Forest, and ANN) [41][42]. These indicators help quantify classifier performance based on various factors, such as specificity, precision, and sensitivity.

International Journal of Scientific Engineering and Science
Volume 9, Issue 10, pp. 16-23, 2025. ISSN (Online): 2456-7361

Vrai Positif (VP or TP): refers to situations in which the model predicts correctly the positive class.
Vrai Negatif (VN or TN): denotes the cases where the model correctly predicts the negative class.
Faux Positif (FP): occurs when the model predicts a positive class while the actual etiquette is negative.
Faux Negatif (FN): refers to situations where the model incorrectly classifies an event as negative because it fails to recognize a positive event.

The following metrics are derived from these four essential elements of the confusion matrix.

2.4.1 Accuracy

The accuracy of a model is a crucial indicator for assessing its effectiveness, especially in classification problems. She illustrates the model's overall ability to make accurate predictions by expressing the percentage of correctly classified examples relative to all evaluated data. This measure is calculated using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2 .4.2 Precision

The precision is a number between 0 and 1 that helps evaluate the reliability of positive predictions made by a classification model. It indicates the percentage of optimistic predictions that are actually correct. Additionally, it provides information on the percentage of true positives among all cases the model has labeled as positive. It is defined by the following formula:

$$Precision = \frac{TP}{TP + FP}$$

2.4.3 Recall

The recall measures the percentage of correctly identified positive cases out of all actual positive cases. Even though mistakes can happen, it is essential when detecting positives is the top priority. It is as follows:

$$Recall = \frac{TP}{TP + FN}$$

2.4.4 Error rate

The actual value is split by the percent value that differs between the actual value and the results that were gathered.

$$Error\ rate = \frac{[Observed\ Value - Actual\ Values]}{Actual\ Values} \text{x } 100$$

2.4.5 F-Measure

The F-measure is also called the F1 score and it is calculated as follows:

$$F - measure = \frac{[2 * Precision * Recall]}{Precision + Recall}$$

III. Materials and Methods

A. Experimental Protocol

Between March and July 2022, a study was conducted at the Quodess Medical Laboratory, located in Sidi Bennour, with the aim of evaluating ABO-Rhesus blood groups while exploring a complementary analysis method based on near-infrared (NIR) spectroscopy. The process began with a secure blood draw performed by a qualified nurse following a strict protocol. Each sample was then divided into two cuvettes containing an anticoagulant (EDTA), ensuring the stability of the blood until analysis. Blood group determination was based on the hemagglutination test (Beth-Vincent test), which involves detecting the presence of antigens on the surface of red blood cells using specific test sera (anti-A, anti-B, and anti-D). The observed agglutination allowed for reliable identification of the ABO blood group and Rhesus factor (Figure 3). In parallel, an experiment was conducted to analyze the same blood samples using NIR spectroscopy, covering a spectral range from 1000 nm to 2600 nm. A drop of blood was placed on a glass slide and exposed to an infrared light source, enabling the recording of transmission spectra. This analysis was performed using two devices: a laboratory-grade spectrometer and a portable NIR sensor connected to a computer via the NeoSpectra interface (Figure 4). The main objective of this approach was to compare the accuracy and reliability of the two systems. The collected data allowed for an assessment of the relevance of the portable NIR sensor for use in clinical or research settings, as a complement to traditional blood typing methods. This integrated approach, combining conventional hemotyping techniques with technological innovation, highlights the growing interest in portable devices for improving diagnostic practices in clinical biology.

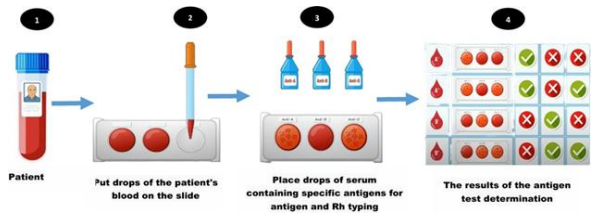

Figure 3: Procedure for ABO and Rh Blood Typing in the Medical Analysis Laboratory ALQODS.

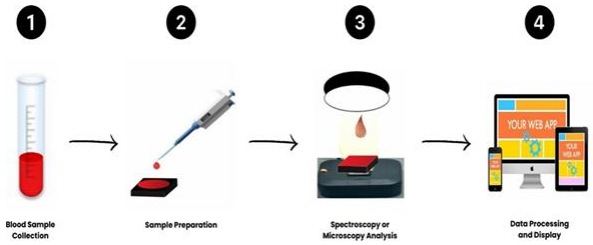

Figure 4: Steps in the Blood Sample Analysis Process Using NIR Spectroscopy

B. Data Collection of ABO Blood Groups

The ABO system was used to type each of the 211 patients in the sample included in this study based on their blood group. Analysis of the relative frequency of each phenotype is made possible by the distribution of patients by blood group, which is crucial for epidemiological research as well as transfusion management. With 110 patients, or 52.1% of the total, Group A has the largest representation in the sample. Fifty of them are Rhesus negative (A–), which accounts for 45.5% of group A, and sixty are Rhesus positive (A+), which accounts for 54.5%. Within group A, this distribution indicates a minor prevalence of Rhesus positive individuals, which is in line with findings in several communities. Thirty-one patients, or 14.7% of the sample, are in Group B. Of them, 30 patients (96.8% of group B) are Rhesus positive (B+), while just one patient (3.2%) is Rhesus negative (B–). The significance of this subgroup, which

is rare and important in blood banks, is highlighted by the high prevalence of Rh-positive and the rarity of B-. With 30 patients, Group AB makes up 14.2% of the sample and exhibits an odd distribution. Ten of these people are AB+ (33.3%), while twenty are AB– (66.7%). Rarely seen in other groups, the overrepresentation of AB– can point to local specificities or sampling bias, requiring more investigation. Regarding group O, the sample included 30 O+ and 10 O– patients, suggesting a high representation of group O, even if specific statistics were not initially detailed. This distribution is in line with worldwide patterns, which show that blood group O is typically more common. Each participant in this study was encoded based on whether antigens A, B, AB, and O were present or absent. An antigen's presence is indicated by a value of 1, and its absence is indicated by a value of 0. This binary system was employed. As a result, blood group A patients were encoded as (A = 1, B = 0, O = 0, AB = 0), blood group O patients as (A = 0, B = 0, O = 1, AB = 0), blood group B patients as (A = 0, B = 1, O = 0, AB = 0), and blood group AB patients as (A = 0, B = 0, O = 0, AB = 1). This encoding makes statistical analysis and categorization easier, enabling quick phenotypic

3.3. Data Preprocessing for Antigen A Detection

Python 3.7 was used to create the preprocessing procedures and the machine learning algorithms, which include RF, SVM, and ANN. Different models were constructed for each method, and the target variable was treated independently. Class imbalance is found in the dataset; group A, which makes up 52.1% of the samples (110 patients), is the majority class. Of them, 50 patients, or 45.5% of group A, are antigen-negative (A–), while 60 patients, or 54.5% of group A, are positive for antigen A (A+). The spectral data of blood droplets in the wavelength region of 1300 to 2600 nm was preprocessed in order to enhance model performance. Important spectral properties were preserved while noise was reduced using a Savitzky- Golay filter, which is frequently used in spectroscopy [1]. Without changing important signal properties, this digital filter efficiently eliminates signal artifacts [2]. The goal of the data preprocessing was to maximize neural networks' capacity for prediction. To assess the models' capacity to identify antigen A in a blood drop, the dataset was divided into training (85%) and testing (15%) sets. In order to guarantee repeatability and preserve a constant sample distribution throughout several experimental runs, the split was carried out pseudo-randomly using a preset seed. To accurately compare how well various models perform on the same datasets, this step is crucial.

3.4. Development and Optimization of Machine Learning Models for Antigen A Detection

This study used customized Python packages to create three machine learning algorithms. Scikit- learn was used to create SVMs and Random Forests [3,4], while Keras [5] and TensorFlow [6] were used to create ANNs. These models were compared according to their overfitting indicators, cross-validation scores, and learning curves.

The Scikit-learn-built SVM model demonstrated high generalization ability with a cross-validation score of 0.98. A discernible discrepancy between training and validation performance, however, indicates that the learning curves indicate a risk of overfitting with limited data quantities (Figure 5). As the sample size grows, this gap tends to close, suggesting good convergence. When the data is normalized and the hyperparameters [7–8] (like the kernel and the regularization parameter C) are adjusted using methods like GridSearchCV, SVM works especially well for compact, high-dimensional datasets. With a training score of 1.00 and a cross- validation score of 0.955 (Figure 6), the Random Forest model performed quite well, indicating only minor overfitting. This model is valued for its low sensitivity to feature scaling and resilience to noisy data. It is advised to raise the minimum number of samples needed to divide a node or restrict the maximum tree depth in order to reduce the risk of overfitting. Random Forest's interpretability is a noteworthy characteristic, especially when it comes to feature importance analysis, which is helpful for choosing the most pertinent predictors.

With a high validation score of 0.987 (Figure 7), the ANN, created with Keras and TensorFlow, performed the best out of the three models. This outcome demonstrates how well the model can represent intricate nonlinear interactions. Neural networks, however, are more likely to overfit and demand more processing power, particularly when the training score is much higher than the validation value. Regularization strategies like Dropout or L2 penalty should be applied to lessen this effect, and learning curves should be regularly observed throughout the training process.

According to the learning curve results, the SVM and Random Forest models perform quite competitively for short datasets, with SVM displaying a marginal validation score advantage. Random Forest offers a great balance of interpretability, performance, and ease of use. In terms of accuracy, the neural network performs better than the other models; however, this comes at the expense of increased computing demand and overfitting risk. Thus, the ultimate decision is contingent upon the particular limitations of every application, such as the complexity of the data, the resources at hand, and the requirement for model interpretability

IV. Results and Discussion

TABLE 1: Performance Results of SVM, Random Forest, and ANN Models in Cross-Validation and Testing.

Model	Phase	Accuracy	Error Rate	Precision	Recall	F-Measure
SVM	Cross-Validation	0.987	0.013	5.8446	0.987	1.689
	Testing	0.986	0.014	_	_	-
RF	Cross-Validation	0.960	0.040	1.5700	0.960	1.191
	Testing	0.950	0.050	_	_	_
ANN	Cross-Validation	0.989	0.011	1.0500	0.989	1.017
	Testing	0.980	0.020	_	_	_

In this study, we compared three machine learning models SVM, RF, and ANN for the classification of antigen A using near-infrared (NIR) spectra collected from a few drops of blood. This problem relies on biological data and involves complex sources of noise, inter-individual variability, and spectral redundancy related to erythrocyte shapes in the blood. The goal was to identify the most appropriate model in terms of accuracy, stability, and generalizability. According to the results presented in Table 2, the SVM model demonstrated high

accuracy in cross-validation testing (0.987) and in actual testing (0.986), with a recall of 0.987, indicating that this model can reliably detect positive cases. However, the reported precision (5.8446) and F-measure (1.689) suggest a calibration issue or a threshold weighting problem—possibly a misbalance in class weighting. This imbalance might be due to the model being overloaded with dominant features, thereby reducing its robustness. Another phenomenon that warrants interpretation is the observed gap in RMSE between testing and validation, suggesting that the model construction process involved inconsistencies or overfitting. On the other hand, the Random Forest model exhibited moderate but stable performance. Cross-validation showed an accuracy of 96%, test validation achieved 95%, with a recall of 0.960 and an F-measure of 1.191. In other words, Random Forest produced more balanced outcomes across the measured metrics. Its performance can be attributed to its structure: as previously mentioned, it consists of multiple independent decision trees. Therefore, the model can mitigate biological variability and the presence of outliers, making it a credible candidate for real-world clinical settings where spectral quality may vary. Finally, the ANN model stood out as the best-performing among the three. With a cross-validation accuracy of 98.9% and 98% in testing, a recall of 0.989, and an F-measure of 1.017, it demonstrates a well-balanced profile and minimal error (1.1%). Its consistency between validation and testing phases reflects a well-regulated learning capacity without signs of overfitting. Moreover, its ability to model complex nonlinear relationships and extract subtle patterns from the NIR spectra gives it a clear advantage in this type of multidimensional problem.

The comparative analysis of the three models reveals complementary profiles. SVM excels at processing data it has already encountered but is more vulnerable to changes not present in the training dataset. Although the absolute accuracy of the Random Forest is lower, its robustness and consistency make it a preferred choice, especially in real clinical situations. As for the artificial neural network, it combines exceptional performance, stability, and generalization ability, making it the most suitable approach for classifying NIR spectra in a biomedical context.

Evaluation on an Independent Dataset and Application Perspectives : To assess the real predictive capacity of the developed models, an independent validation phase was carried out on a new set of 30 blood samples, distinct from those used during the training and cross-validation phases. This set included 20 women and 10 men, representing a moderate level of biological diversity within the tested population. Among these samples, 15 were identified as positive for antigen A, while the remaining 15 belonged to blood groups that did not present this antigen. The Artificial Neural Network (ANN) model, applied to these new NIR spectra, achieved a prediction score of 0.89, confirming its ability to effectively detect the presence of antigen A, even on entirely new and previously unseen data. Although this score is slightly lower than the performance recorded during cross-validation and testing (≈98%), it remains very encouraging given the unknown nature of the data and the inter- individual variability. These results demonstrate that the ANN model exhibits a strong generalization ability beyond the training distribution, which is essential for any real-world clinical application. With a view toward integration in a laboratory setting, particularly within ALQODS Laboratory, this model can be calibrated and optimized using internal procedures or an automated adjustment system, enabling dynamic adaptation of decision thresholds based on the specific characteristics of the biological samples being analyzed. Such calibration is expected to raise the model's accuracy to levels approaching 99%, thus meeting the standards of reliability, sensitivity, and reproducibility required for non-invasive blood typing in biomedical contexts.

V. Conclusion and Future Directions

This study demonstrates the feasibility and effectiveness of applying machine learning models, specifically SVM, RF, and ANN to the classification of blood groups through near-infrared (NIR) spectral data. Focusing on the detection of antigen A, we proposed and evaluated a pipeline that integrates spectral data acquisition, feature processing, and model training and validation using a well-defined and replicable methodology. The ANN model outperformed its counterparts across most performance metrics, achieving up to 98.9% accuracy during cross-validation and maintaining strong generalization on independent testing data. Moreover, in an additional validation phase using 30 new samples (including 20 females and 10 males), the ANN achieved a prediction score of 0.89, accurately identifying 15 true positives, thus demonstrating its practical potential for real-world biomedical deployment. Compared to prior work, this study introduces a novel integration of NIR spectroscopy with modern machine learning classifiers, in a minimally invasive setting, using a compact data acquisition setup and a relatively small sample volume (one drop of blood). It further validates its findings through quantitative comparisons, statistical metrics, and independent test data, ensuring methodological soundness and reproducibility. However, several limitations must be acknowledged. First, the current sample size, while sufficient for initial validation, remains limited in terms of population variability (e.g., age, ethnicity, health conditions). Second, although the ANN showed strong generalization, further calibration is required to enhance its decision threshold stability when transitioning from controlled experimental conditions to clinical practice. Third, the data relied solely on spectral patterns, without incorporating complementary biomarkers or demographic metadata, which could further improve prediction accuracy. Despite these limitations, the results are promising and pave the way for the development of non-invasive, low-cost, and rapid blood typing tools, especially in resource- limited clinical settings. The successful integration of machine learning and NIR spectroscopy opens new research opportunities in biomedical diagnostics and personalized healthcare.

VI. Future Research Directions

Future studies will focus on:

A. Scaling up the dataset to include more diverse and balanced population groups,

B. Incorporating additional blood group antigens (e.g., B, AB, O, Rh) to expand classification capability,

C. Integrating complementary physiological or biochemical features with NIR spectra for multimodal analysis,
D. Developing a real-time prediction system deployable in clinical laboratories like ALQODS,
E. Investigating more sophisticated deep learning architectures for improved pattern identification and robustness, such as convolutional or recurrent neural networks.

Ultimately, this work contributes to the advancement of AI-driven diagnostics and supports the broader vision of automated, accessible, and accurate blood analysis using portable and non-invasive technologies.

REFERENCES

[1] Goodnough, L. T., & Panigrahi, A. K. (2017). Blood transfusion therapy. *Medical Clinics*, *101*(2), 431-447.
[2] Thein, S. L., Pirenne, F., Fasano, R. M., Habibi,A., Bartolucci, P., Chonat, S., ... & Stowell, S. R. (2020). Hemolytic transfusion reactions in sickle cell disease: underappreciated and potentially fatal. *Haematologica*, *105*(3), 539.
[3] Cutbush, M., Mollison, P. L., & Parkin, D. M. (1950). A new human blood group. *Nature*, *165*(4188), 188-189.
[4] Högman, C. F., & Meryman, H. T. (1999). Storage parameters affecting red blood cell survival and function after transfusion. *Transfusion medicine reviews*, *13*(4), 275-296.
[5] Shim, H., Hwang, J. H., Kang, S. J., Seo, H. S.,Park, E. Y., Park, K. U., & Kong, S. Y. (2020).Comparison of ABO isoagglutinin titres by three different methods: tube haemagglutination, micro- column agglutination and automated immunohematology analyzer based on erythrocyte-magnetized technology. *Vox Sanguinis*, *115*(3), 233-240.
[6] Flegel, W. A. (2011). Molecular genetics and clinical applications for RH. *Transfusion and Apheresis Science*, *44*(1), 81-91.
[7] Anstee, D. J. (2009). Red cell genotyping and the future of pretransfusion testing. *Blood, The Journal of the American Society of Hematology*, *114*(2), 248- 256.
[8] Fichou, Y., Audrézet, M. P., Guéguen, P., Le Maréchal, C., & Férec, C. (2014). Next-generation sequencing is a credible strategy for blood group genotyping. *British journal of haematology*, *167*(4), 554-562.
[9] Storry, J. R., & Olsson, M. L. (2009). The ABO blood group system revisited: a review and update. *Immunohematology*, *25*(2), 48.
[10] Hosseini-Maaf, B., Hellberg, Å., Rodrigues, M. J., Chester, M. A., & Olsson, M. L. (2003). ABO exon and intron analysis in individuals with the A weak B phenotype reveals a novel O 1 v-A 2 hybrid allele that causes four missense mutations in the A transferase. *BMC genetics*, *4*, 1-11.
[11] Kappler-Gratias, S., Peyrard, T., Rouger, P., Le Pennec, P. Y., & Pham, B. N. (2010). Blood group genotyping by high-throughput DNA analysis: Application to the French panel of RBC reagents. *Transfusion clinique et biologique*, *17*(3), 165-167.
[12] Dzik, W. H. (2003). Emily Cooley Lecture 2002: transfusion safety in the hospital. *Transfusion*, *43*(9), 1190-1199.
[13] Baker, M. J., Byrne, H. J., Chalmers, J., Gardner, P., Goodacre, R., Henderson, A., ... & Sulé- Suso, J. (2018). Clinical applications of infrared and Raman spectroscopy: state of play and future challenges. *Analyst*, *143*(8), 1735-1757.
[14] Fernández-González, A., Obaya, Á. J., Chimeno-Trinchet, C., Fontanil, T., & Badía-Laíño, R. (2023). Viability of ABO Blood Typing with ATR-FTIR Spectroscopy. *Applied Sciences*, *13*(17), 9650.
[15] De Niz, M., Pereira, S. S., Kirchenbuechler, D., Lemgruber, L., & Arvanitis, C. (2025). Artificial intelligence-powered microscopy: Transforming the landscape of parasitology. *Journal of Microscopy*.
[16] Ogunlade, B., Tadesse, L. F., Li, H., Vu, N., Banaei, N., Barczak, A. K., ... & Dionne, J. A. (2024). Rapid, antibiotic incubation-free determination of tuberculosis drug resistance using machine learning and Raman spectroscopy. *Proceedings of the National Academy of Sciences*, *121*(25), e2315670121.
[17] Priyadharshini, K., Mathias, A., Krishnan, R. S., Kanthimathi, N., Raj, J. R. F., & Malar, P. S. R. (2025, April). A Multi-Resolution Deep Learning Approach for IoT-Integrated Biomedical Signal Processing using CNN-LSTM. In *2025 5th International Conference on Trends in Material Science and Inventive Materials (ICTMIM)* (pp. 1362-1369). IEEE.
[18] Liu, G., Wu, Y., Wang, Y., Ye, W., Wu, M., & Liu, Q. (2024). Smartphone-Based Portable Sensing Systems for Point-of-Care Detections. *Portable and Wearable Sensing Systems: Techniques, Fabrication, and Biochemical Detection*, 89-110.
[19] Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. *Current genomics*, *22*(4), 291-300.
[20] Cyran, K. A., Kawulok, J., Kawulok, M., Stawarz, M., Michalak, M., Pietrowska, M., ... & Polańska, J. (2013). Support vector machines in biomedical and biometrical applications. In *Emerging paradigms in machine learning* (pp. 379-417). Berlin, Heidelberg: Springer Berlin Heidelberg.
[21] Cyran, K. A., Kawulok, J., Kawulok, M., Stawarz, M., Michalak, M., Pietrowska, M., ... & Polańska, J. (2013). Support vector machines in biomedical and biometrical applications. In *Emerging paradigms in machine learning* (pp. 379-417). Berlin, Heidelberg: Springer Berlin Heidelberg.
[22] Nayak, R., Jain, L. C., & Ting, B. K. H. (2001).Artificial neural networks in biomedical engineering: a review. *Computational Mechanics– New Frontiers for the New Millennium*, 887-892.
[23] Chowdhuri, S., Bhattacharjee, M., Bhowmick, A., Ghosh, S., Das, S., Garika, N. K., ... & Pyne, A. (2023). Heart Disease Diagnosis Using Machine Learning Algorithms. In *Renewable Resources and Energy Management* (pp. 403-410). CRC Press.
[24] Gargeya, R., & Leng, T. (2017). Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, *124*(7), 962-969.
[25] Zubair, M., Tripathy, R. K., Alhartomi, M., Alzahrani, S., & Ahamed, S. R. (2023). Detection of Sleep Apnea From ECG Signals Using Sliding Singular Spectrum Based Subpattern Principal Component Analysis. *IEEE Transactions on Artificial Intelligence*, *5*(6), 2897-2906.
[26] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.
[27] Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.
[28] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121-167.
[29] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*, 199-222.
[30] Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning. Springer series in statistics. *New York, NY, USA*.
[31] Wang, S., Zhang, J., Fu, Y., & Li, Y. (2011).ACM transactions on intelligent systems and technology. In *ACM Transactions on Intelligent Systems and Technology*.
[32] Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.
[33] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
[34] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18- 22.
[35] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.
[36] Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? The journal of machine learning research, 15(1), 3133-3181.
[37] Heaton, J. (2018). Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618. Genetic programming and evolvable machines, 19(1), 305-307.
[38] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.
[39] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.
[40] Aurélien, G. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. o'reilly.

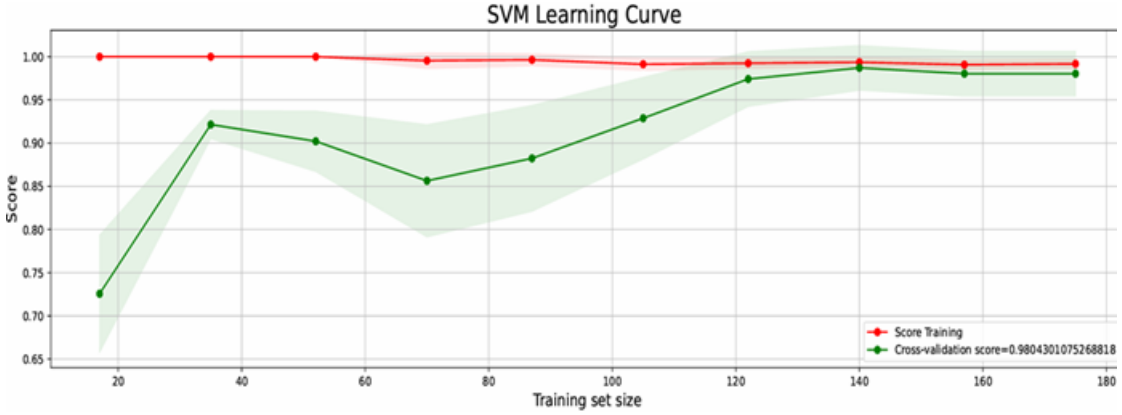

Figure 5: SVM Learning Curve

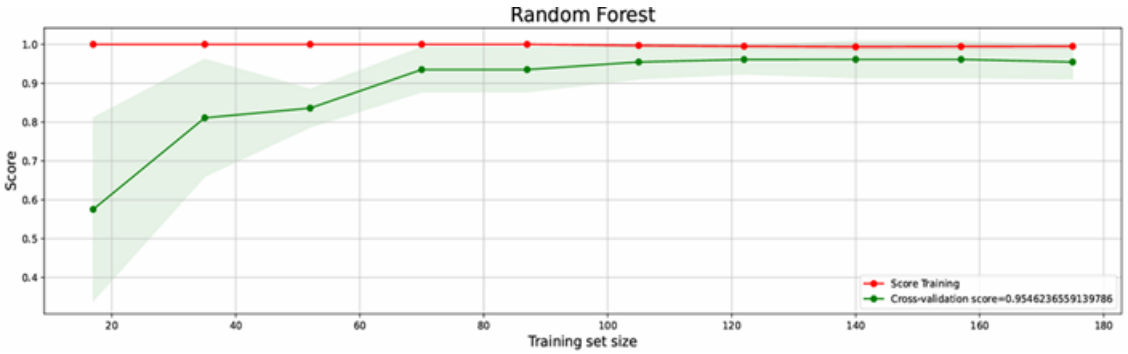

Figure 6: Random Forest Learning Curve

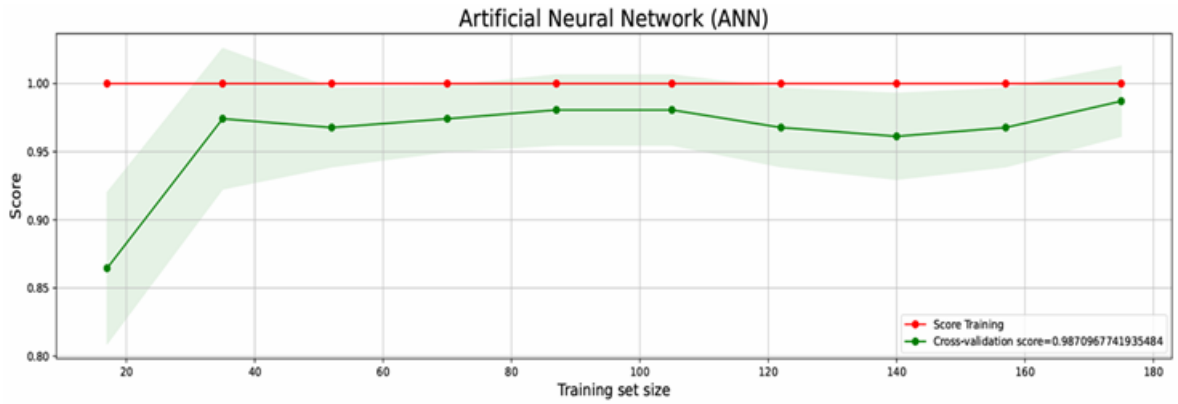

Figure 7: Artificial Neural Network (ANN) Learning Curve