Expert-level detection of M-proteins in serum protein electrophoresis using machine learning

Eike Elfert[1†], Wolfgang E. Kaminski[1,2†], Christian Matek[3], Gregor Hoermann[4], Eyvind W. Axelsen[5,6], Carsten Marr[3‡], Armin P. Piehler[*4,5‡]

Corresponding author:

Armin P. Piehler, MD PhD,

armin.piehler@mll.com

[1]University of Heidelberg, Medical Faculty Mannheim, 68167 Mannheim, Germany;

[2]ingenium digital diagnostics GmbH, 60314 Frankfurt a.M, Germany;

[3]Institute of AI for Health, Helmholtz Munich -

German Research Center for Environmental Health, 85764 Neuherberg, Germany;

[4]MLL Munich Leukemia Laboratory, 81377 Munich, Germany;

[5]Fürst Medical Laboratory, 1051 Oslo, Norway;

[6]University of Oslo, Department of Informatics, 0313 Oslo, Norway.

*Address correspondence to these author at:

Armin P. Piehler, MD PhD

Munich Leukemia Laboratory GmbH,

Max-Lebsche-Platz 31, 81377 Munich, Germany.

Fax/Tel: +49 89 99017-0/ +49 89 99017-111.

E-mail: armin.piehler@mll.com.

Running head/short title: M-protein detection using machine learning

M-protein detection using machine learning

[†]Eike Elfert and Wolfgang E. Kaminski contributed equally to this work.

[‡]Carsten Marr and Armin P. Piehler contributed equally to this work.

Word count abstract: 244

Word count article: 3492

Number of tables: 3

Number of figures: 5

Supplemental tables: 1

M-protein detection using machine learning

Nonstandard abbreviations:

SPE, serum protein electrophoresis;

IMT, immunotyping;

M-proteins, monoclonal proteins;

ML, machine learning;

CNN, convolutional neural network;

RFC, random forest classifier;

MGUS, Monoclonal gammopathy of undetermined significance;

IFE, immunofixation electrophoresis;

AUC, area under the curve;

ROC, receiver operating characteristic;

Abstract

Objectives

Serum protein electrophoresis (SPE) in combination with immunotyping (IMT) is the diagnostic standard for detecting monoclonal proteins (M-proteins). However, interpretation of SPE and IMT is weakly standardized, time consuming and investigator dependent. Here, we present five machine learning (ML) approaches for automated detection of M-proteins on SPE on an unprecedented large and well-curated data set and compare the performance with that of laboratory experts.

Methods

SPE and IMT were performed in serum samples from 69 722 individuals from Norway. IMT results were used to label the samples as M-protein present (positive, n=4273) or absent (negative n=65 449). Four feature-based ML algorithms and one convolutional neural network (CNN) were trained on 68 722 randomly selected SPE patterns to detect M-proteins. Algorithm performance was compared to that of an expert group of clinical pathologists and laboratory technicians (n=10) on a test set of 1 000 samples.

Results

The random forest classifier showed the best performance (F1-Score 93.2%, accuracy 99.1%, sensitivity 89.9%, specificity 99.8%, positive predictive value 96.9%, negative predictive value 99.3%) and outperformed the experts (F1-Score 61.2 ± 16.0%, sensitivity 94.3 ± 2.8%, specificity 88.9 ± 10.9%, positive predictive value 47.3 ± 16.2%, negative predictive value 99.5 ± 0.2%) on the test set. Interestingly the performance of the RFC saturated, the CNN performance increased steadily within our training set (n=68 722).

Conclusion

Feature-based ML systems are capable of automated detection of M-proteins on SPE beyond expert-level and show potential for use in the clinical laboratory.

Keywords: artificial intelligence; monoclonal gammopathy; myeloma; electrophoresis; machine learning

Introduction

Monoclonal proteins (M-proteins) are an overproduction of a single immunoglobulin or a fragment of it and represent the hallmark feature of plasma cell dyscrasia [1]. Plasma cell dyscrasias encompass a wide variety of entities from monoclonal gammopathy of undetermined significance (MGUS), a pre-malignancy with prevalences of > 3% (>5%) in patients >50 (>70) years of age [1], to malignant entities including multiple myeloma, Waldenstrom's macroglobulinemia, primary AL amyloidosis, and plasma cell leukemia. Multiple myeloma is the third most common hematologic malignancy and responsible for about 0.9% of all new cancer diseases worldwide [2]. Progression from MGUS to malignant entities is usually accompanied by an increase of M-protein concentration [1].

Immunofixation electrophoresis (IFE) is the current gold standard for detection of M-proteins. Both, IFE and immunotyping (IMT), are methods used to establish the monoclonal origin of suspected monoclonal proteins and to characterize their isotype and light chain composition [3–8]. Serum protein electrophoresis (SPE) is significantly less sensitive and less specific in detecting M-proteins than IFE or IMT. Still, it is widely used as a first-line detection system for M-proteins as it is a robust, analytically straightforward and inexpensive method [3–5].

In SPE, results are presented as a continuous graph based on the densitometric analysis of the electrophoretically separated serum protein pool [9]. Monoclonal gammopathies often present as a clearly delineated, discrete M-protein peak in the SPE curve, which is readily detectable by visual inspection (Figure 1A) compared to an SPE curve without M-protein peak (Figure 1B). Frequently, however, the monoclonal protein may be difficult to distinguish from other abnormalities that manifest in SPE and may be misinterpreted or escape identification (Figure 1C) [9]. Moreover, individual expertise in evaluating SPE curves lead to great variability of classification performance and may be biased by respective specialist focus that adds another layer of complexity to the interpretation of SPE patterns [10–13].

Machine learning (ML) algorithms have recently been introduced to perform classification tasks in diagnostic settings [12–17]. In the clinical laboratory, only sporadic efforts have been made during the past two decades to establish automated analysis of SPE patterns [18–22]. The patient cohort sizes in most of these studies were limited or sample labeling relied solely on the subjective SPE pattern analysis performed by laboratory technicians or hematologists. In none of these reports the presence of M-proteins has been confirmed by IMT [18–22].

In the present study, we trained, optimized and evaluated five ML algorithms that classify SPE patterns into M-protein positive or negative. The data set comprises nearly 70 000 individual SPEs that were manually annotated using IMT results as ground truth. Furthermore, we demonstrate that the most effective M-protein classifier outperforms experienced laboratory staff.

Materials & Methods

Data set

Serum protein electrophoresis (SPE) and immunotyping (IMT) were performed on Capillarys 2 Flex Piercing instruments ® (SEBIA, Lisses, France) according to the manufacturer's instructions. In the study period, more than 500,000 SPEs were analyzed, and IMT was added when clinical information given by the requesting physician indicated the presence of an M-peak or monoclonal gammopathy. Moreover, all SPEs were assessed by a laboratory technician and, in case of obvious M-spikes or irregularities in alpha-2 through gamma fractions, samples were further processed by IMT. Each IMT was manually evaluated by visual inspection of a clinical pathologist and labeled as M-protein present (positive) or absent (negative). Only samples in which both SPE and IMT were performed, and only one sample of each individual were included in the data set. SPE recorded at Fürst Medical Laboratory, Oslo, during 2007-2021 were extracted. The SPE graph of each sample contains 300 data points of density levels measured by absorbance photometry.

Patient data were anonymized prior to further analysis. The study was approved by the Ethics Committee of the University of Heidelberg (2018-548N-MA), Germany, and the Regional Ethics Committee REC South-East, Norway (231395).

Training/test-split

The samples were divided into a training set of 68 722 samples and a test set of 1000 samples. A 10-fold cross validation was performed on the training set [23]. Each of the ten training subsets comprised 58 067 IMT negative and 3783 IMT positive samples. Each validation set included 6451 IMT negative and 421 IMT positive samples. The best performing algorithm was then trained on the complete training set (n=68 722).

The test set of 1000 samples was randomly selected and reserved for performance comparison with expert laboratory technicians and physicians. It included 931 IMT negative and 69 IMT positive samples. The ratio of IMT positives to negatives was about 1:14 and comparable in the training and test set.

Machine Learning

After a pre-selection of machine learning algorithms for classification problems offered by Scikit-learn [23, 24], the algorithms presented in the current study performed superior to the rest. Moreover, deep learning algorithms have shown to outperform classical algorithms in some classification tasks, especially in image classification, and therefore a convolutional neural network (CNN) was included as well [25, 26].

The following five ML algorithms were used: random forest classifier (RFC), extremely randomized trees classifier, adaboost classifier, gradient boosting classifier and a convolutional neural network (CNN) [27–31] based on the structure reported by Liu et al. [32] using ReLU as activation function. Our CNN consisted of eighteen layers with three convolutional layers and a total number of 194 945 parameters. All classifiers were taken from the python library scikit-learn [23, 24] and keras [33, 34], respectively. All results are reported as mean ± standard deviation.

SPE classification by laboratory experts

To compare the performance of M-protein detection between human and machine, ten experienced clinical pathologists and laboratory technicians, who perform the daily routine evaluation of SPE, were recruited from Fürst Medical Laboratory and MLL Munich Leukemia Laboratory. The panel of experts had on average 17.4 ± 13.5 (mean ± standard deviation) years of professional expertise and 8.0 ± 7.3 years of experience in classifying SPE. During this period of time, they had classified a total number of 34 400 ± 32 500 SPEs. Each individual expert had the task to decide whether or not to recommend IMT based on the suspicion of the presence of a M-peak on the test set, which contained a total of 1000 randomly selected SPE profiles, presented by a computer program. The program sequentially displayed the graphs of the SPEs and the user was requested to click on the 'recommend IMT' or 'do not recommend IMT' option. There was no time limit, and the experts had the option to go back and forth in the sequence of the presented SPE graphs and to change their decision. The average time to complete classification of the entire test set was 26.0 ± 11.3 minutes.

Performance evaluation of algorithms

Performance of the algorithms was assessed by calculating accuracy, sensitivity, specificity, negative and positive predictive value. The mean ROC-AUC over a 10-fold cross-validation was utilized to compare the performance between the different ML algorithms. For comparison of the performance of the experts with that of the best performing algorithm the receiver operating characteristic (ROC) [35] curve and the F1-Score was used.

Results

Data set compilation

We assembled a data set of 69 722 SPE/IMT pairs, generated at Fürst Medical Laboratory during routine testing. The data set contained 4273 (6.1%) IMT positive and 65 449 (93.9%) IMT negative samples (Figure 1D). The median age of the patients was 58 years ranging from 0 to 103 years and an interquartile range of 29 (Figure 1E). With IMT as ground truth, the classification task was to distinguish IMT positive from IMT negative samples based on the individual SPE profiles (Figure 1A, B). Of note, the IMT positive data set also included low concentrated M-proteins that are particularly difficult to differentiate from other causes of minor SPE profile abnormalities (Figure 1C).

Data were randomly subdivided into a training set (n=68 722 samples) for optimizing ML methods and a test set (n=1000 samples) for comparing machine versus human performance.

Feature extraction and engineering

For classical ML algorithms, high quality features are essential [36]. To identify them, expert laboratory technicians and physicians (n=10) were interviewed to indicate crucial factors for evaluating SPE patterns. As a result, the following 270 features were designed and used in the study: (i) the x and y-coordinates of the first five peaks in the β- and γ-globulin fractions (Figure 2A,B); (ii) the partial AUCs around the peaks calculated applying Simpson’s rule [37] (Figure 2C), and (iii) the gradients around the peaks calculated between two adjacent x- and y-coordinates in the β- and γ-globulin fractions (Figure 2D). From that list, 13 AUCs and gradients prior and after the x-coordinate of the five peaks were chosen (Figure 2C and 2D). In order to have equally long examination intervals the start of the β-globulin fraction was set at x-coordinate 165 in each SPE (Figure 2B). When an SPE graph displayed less than five peaks or when the 13 AUCs or gradients in proximity of the peaks were out of the data range, those values were set to zero.

Evaluation of machine learning algorithms on the training set

The performances of the different ML algorithms on the training sets are shown in Table 1. Overall, ROC-AUC, accuracy, sensitivity, specificity, positive and negative predictive value were > 78%. The random forest classifier (RFC) showed best performance with an ROC-AUC of 92.7 ± 0.8% and sensitivity of 85.7 ± 1.5%. Compared to the CNN, the RFC performed better in three out of six measures (ROC AUC, sensitivity and positive predictive value), whereas the CNN showed a slightly higher performance in accuracy and negative predictive value.

[Place Table 1 near here]

Most important features

In contrast to earlier studies [19–22], our approach is interpretable. We can thus determine feature importance for the RFC by calculating the impurity decrease for each feature [23, 24]. We find that the top ten features all describe the shape of the third peak (Supplemental table 1). This fits to expert knowledge: Usually there are only three peaks in the examined spectrum, the β1-, the β2- and the γ-peak (Figure 2A). Most M-gradients occur in the γ-fraction, which represents the third peak.

The random forest classifier outperforms experienced laboratory staff on the test set

As the best performing algorithm, RFC was trained using the complete training set comprising 68 722 samples. The resulting model was then evaluated on the test set of 1000 samples that was reserved for comparing human and computer performance. The test set contained 931 IMT negative and 69 IMT positive samples. The results of the experts and the random forest classifier with a prediction probability threshold of 0.5 on the sample test set are shown in Table 2.

The RFC model correctly classified 929 SPE with negative IMT as ‘M-protein absent’, and 62 SPE with positive IMT as ‘M-protein present’ (Figure 3A). Nine cases were misclassified, seven false negative and two false positive cases (Figure 3A). On the test set, the RFC reached an F1-score of 93,2% with a sensitivity of 89,9%, a specificity of 99.8%, positive predictive value of 96.9% and a negative predictive value of 99.3% (Table 2).

[Place Table 2 near here]

In contrast, the mean performance of a panel of ten laboratory experts on the sample test set were as follows: F1-score of 61.2 ± 16.0%, sensitivity was 94,3 ± 2,8%, specificity 88,9 ± 10,9%, positive predictive value of 47.3 ± 16.2% and a negative predictive value of 99.5 ± 0.2 %.

Thus, the RFC proved superior to the average of the experts in four of the six performance categories (Table 2).

Adjustment of algorithm performance

ROC curve analysis revealed that the RFC algorithm yielded mostly higher sensitivity and specificity scores than the individual experts (Figure 3B). Using the prediction probability threshold to adjust the sensitivity and specificity to each individual's performance demonstrated that only two of the ten experts performed better than the random forest classifier. The laboratory technician with the highest sensitivity score (100.0%) exhibited a test specificity of 88.3%, whereas the RFC algorithm showed a specificity of 0.0% when adjusted to the same sensitivity. The technician with the highest specificity (96.6%) achieved a sensitivity of 91.3%. Under these conditions the RFC algorithm yielded a sensitivity of 94.2%.

Altering the algorithm's prediction probability threshold to a value of 0.4 corresponding to a sensitivity close to the average of the laboratory experts results in the following performance measures of the RFC: 94.2% sensitivity, 99.7% specificity, 95.6% positive predictive value, 99.6% negative predictive value (Table 3).

[Place Table 3 near here]

In addition, we trained the algorithm on a balanced training data set to improve sensitivity. By reducing IMT negative samples to an IMT positive to IMT negative ratio of 1:1 in the training set, a newly trained RFC shows an increased sensitivity of 94.2% and specificity of 96.9% on the test set.

Comparison of the performance between classical and deep learning algorithms

Deep learning algorithms have regularly shown to outperform classical feature-based methods when available training data sets are large [25, 26]. In contrast to these studies, the CNN approach was inferior to the RFC in three out of six performance categories in the present work (Table 1). Therefore, we tested the effect of varying training set sizes from 1 to 100% in steps of 687 (1%), 1374 (2%), 3436 (5%), 6872 (10%), 13 744 (20%), 34 361 (50%) and 68 722 (100%) on the performance of both algorithms. The performance of the RFC algorithm reached a plateau at about 10% of the complete training data set (n=6 872 SPEs) (Figure 4). In contrast, the performance of the CNN showed no saturation even when trained on the complete training set (Figure 4).

Discussion

Largest dataset with ground truth

In the present study, we trained five different ML algorithms to predict the presence or absence of an M-protein evaluating serum protein electrophoresis. Compared to previously published studies with a significant lower number of cases or the lack of access to immunofixation electrophoresis (IFE) or immunotyping (IMT) results as ground truth [18–22], we used a high-quality data set of almost 70 000 SPEs annotated for the presence or absence of an M-protein by IMT**.** We identified a random forest classifier (RFC) as the best-performing algorithm and demonstrated that it achieves accuracy levels exceeding that of human experts for classification of M-protein presence.

Compared to IMT with a detection limit of about 0.25 g/L, IFE shows a higher sensitivity of identifying low-concentration M-peaks with a detection limit 0.1 g/L [7, 8], and thus some M-proteins with low concentration may be missed in our study. However, both IFE and IMT are acknowledged methods to detect and further characterize M-proteins after screening with SPE, which has significantly lower sensitivity and specificity [6].

Advantages

Our best-performing RFC algorithm provides several advantages over conventional SPE pattern evaluation in the clinical laboratory. First, AUC of ROC curve analyses demonstrate that it consistently performs at least comparable to highly experienced experts with a professional record of > 30 000 SPE pattern classifications (Figure 3B). Second, the RFC algorithm runs with less variability relative to laboratory technicians and physicians who exhibited high inter-individual variability. This problem of variance of classification performance can be seen in other scientific works as well [10–13]. Third, RFC-based M-protein classification is 1000-fold faster (26.0 ± 11.3 minutes in contrast to less than 1 second) and can run 24/7 making it accessible even outside normal working hours. Fourth, the trained algorithm analyzes each individual SPE pattern on the same level of accuracy and robustness and human factors such as operator fatigue or distractibility [38] are eliminated a priori. Finally, unlike in human operators, our trained RFC algorithm allows flexible adjustment of the test sensitivity and specificity depending on the clinical context. Taken together, our results demonstrate that a trained ML algorithm can perform automated detection of M-proteins on SPE with accuracies exceeding that of experts. The RFC-based "M-protein identifier" is highly efficient and can readily be integrated into the workflow of a routine laboratory as a digital decision support system for laboratory experts (Figure 5). Its unprecedented speed enables the accurate analysis of hundreds to thousands of SPEs within seconds, and it may thus be particularly attractive to medium and high-throughput laboratories.

ML algorithms still lack the ability to integrate all information such as clinical context to really make a profound and intelligent decision. The presented ML algorithm is not intended to replace human operators, but to improve detection of M-peaks by functioning as a diagnostic supporting tool.

The results are validated on internal data only. Further studies are needed to confirm the transferability on external data.

Superior feature-based model

Frequently, tree-based models show a better performance in classification tasks on tabular data compared to deep learning algorithms [39]. This was also observed in the present study in which the RFC outperformed the CNN. The RFC performance already saturated after training with only a small part (about 10%) of the full data set. The CNN, however, showed no clear saturation of performance with increasing data size. Future studies on even larger data sets may exhibit whether prediction accuracy can be further improved and whether CNN performance increases with larger data sets.

Previous studies

Some studies have previously reported the use of mathematical classification and ML algorithms for detection of monoclonal proteins in SPE [18–22]. In all these studies, classification was solely based on SPE patterns and not on the highly sensitive method of IFE or IMT. Numerous ML studies have demonstrated that the quality and quantity of the training data is a critical determinant of the performance of an algorithm [25, 26]. Previous algorithm based SPE studies relied mostly on sample sizes of about 100-5000 SPE curves [18, 19, 21]. Our data set of almost 70 000 individual patient samples exceeds these studies by an order of magnitude. In the study by Chabrun et al [22], machine learning models were trained on 150 000 samples. However, this study employed human experts to establish the ground truth and not IFE or IMT. In our study, results from immunotyping [6] whereas used as ground truth. Interestingly, a recent study employed deep learning approaches to automatically interpret immunofixation with convincing results [40]. In contrast to earlier studies [19–22], our approach is interpretable. We were able to show the most important features for the RFC algorithm (Supplemental table 1).

Of 1000 test samples, our algorithm misclassified only nine SPE samples, seven false negative and two false positive. All seven false negative cases consistently represented SPEs without clearly visible M-protein peaks, probably due to low serum concentration of the monoclonal protein. The two false positive cases were SPEs with irregularities in the β- and γ-fractions (Figure 3A) possibly caused by oligo-/ polyclonal immunoglobulin production due to an inflammatory response. Reinspection of these cases revealed no obvious reason for the misclassification. In addition, erroneous interpretation of the IMTs representing the ground truth in some of these cases cannot be ruled out completely.

The test set samples were selected randomly and the ratio of positive and negative cases was kept constant compared to the training set. Therefore, not all kind of M-proteins that are difficult to detect, e.g. low-concentrated M-proteins with polyclonal background or M-proteins superimposed to other proteins, might be represented in the test set.

Future studies

In machine learning classification tasks, the algorithm calculates individual probabilities of an instance for the respective classes. Subsequently, classification of single events is performed by applying a pre-defined threshold value. Altering this threshold value enables the adjustment of the test measures like sensitivity and specificity (Table 3). This fact can be used to fit the algorithm performance to the pre-test probability of an event or disease. In our case, when aiming for a high sensitivity (i.e. achieving a low number of false negatives), adjustment of the threshold of the trained RFC algorithm resulted in an adjusted sensitivity of 98.6% with a specificity of 76.4% on the test set (Table 3). As another approach to improve sensitivity, a balancing of the training data to an equal number of M-protein positive and negative samples resulted in an adjusted sensitivity of 94.2% with a specificity of 96.9% of the algorithm. Using these approaches the algorithm can be optimized for scenarios such as general population screening or follow-up testing of multiple myeloma patients under treatment. However, both approaches were performed retrospectively, and thus further studies are needed to evaluate test performance in independent data sets.

In this context, it also needs to be pointed out that trained ML algorithms, unlike classical rule-based systems, contain numerous cryptic elements and thus behave to a certain degree like a “black box” [41, 42]. This important fact highlights the need for permanent supervision of such modern classifier systems by medical professionals and precludes their current use as autonomous systems in clinical decision making [41, 42]. Currently, a lot of effort is put into explainable AI and developing ways to decipher and illustrate the decision process of AI-based decision systems [41, 42].

Funding and Disclosures/Conflict of Interest

C.M. got funding from the European Research Council (ERC), grant agreement no. 866411 and support from the Hightech Agenda Bayern.

Acknowledgments

We would like to thank the laboratory technicians at Fürst Medical Laboratory for their participation in the study and their interpretation of protein electrophoresis curves in order to compare performance of the machine learning model with human experts. C.M. acknowledges funding from the European Research Council (ERC), grant agreement no. 866411 and 101113551.

References

1. Kyle RA, Larson DR, Therneau TM, Dispenzieri A, Kumar S, Cerhan JR, et al. Long-Term Follow-up of Monoclonal Gammopathy of Undetermined Significance. N Engl J Med 2018;378:241–9.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424.
3. Rajkumar SV, Vincent Rajkumar S, Dimopoulos MA, Palumbo A, Blade J, Merlini G, et al. International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. Lancet Oncol 2014;15:538-48.
4. Tate JR. The Paraprotein - an Enduring Biomarker. Clin Biochem Rev 2019;40:5–22.
5. Harris NS, Winter WE. Multiple Myeloma and Related Serum Protein Disorders: An Electrophoretic Guide. Demos Medical Publishing 2012.
6. Thoren KL, McCash SI, Murata K. Immunotyping Provides Equivalent Results to Immunofixation in a Population with a High Prevalence of Monoclonal Gammopathies. J Appl Lab Med 2021;6:1551–60.
7. Cárdenas MC, García-Sanz R, Puig N, Pérez-Surribas D, Flores-Montero J, Ortiz-Espejo M, et al. Recommendations for the study of monoclonal gammopathies in the clinical laboratory. A consensus of the Spanish Society of Laboratory Medicine and the Spanish Society of Hematology and Hemotherapy. Part I: Update on laboratory tests for the study of monoclonal gammopathies. Clinical Chemistry and Laboratory Medicine (CCLM). 2023;61:2115–30.
8. Cárdenas MC, García-Sanz R, Puig N, Pérez-Surribas D, Flores-Montero J, Ortiz-Espejo M, et al. Recommendations for the study of monoclonal

gammopathies in the clinical laboratory. A consensus of the Spanish Society of Laboratory Medicine and the Spanish Society of Hematology and Hemotherapy. Part II: Methodological and clinical recommendations for the diagnosis and follow-up of monoclonal gammopathies. Clinical Chemistry and Laboratory Medicine (CCLM). 2023;61:2131–42.

9. O'Connell TX, Horita TJ, Kasravi B. Understanding and interpreting serum protein electrophoresis. Am Fam Physician 2005;71:105–12.
10. Font P, Loscertales J, Soto C, Ricard P, Novas CM-, Martín-Clavero E, et al. Interobserver variance in myelodysplastic syndromes with less than 5 % bone marrow blasts: unilineage vs. multilineage dysplasia and reproducibility of the threshold of 2 % blasts. Ann Hematol 2015;94:565–73.
11. Fuentes-Arderiu X, Dot-Bach D. Measurement uncertainty in manual differential leukocyte counting. Clin Chem Lab Med 2009;47:112–5.
12. Matek C, Schwarz S, Spiekermann K, Marr C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. Nat Mach Intell 2019;1:538–44.
13. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.
14. Shalev-Shwartz S, Ben-David S. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press 2014.
15. Bizopoulos P, Koutsouris D. Deep Learning in Cardiology. IEEE Rev Biomed Eng 2019;12:168–93.
16. McBee MP, Awan OA, Colucci AT, Ghobadi CW, Kadom N, Kansagra AP, et al. Deep Learning in Radiology. Acad Radiol 2018;25:1472–80.

17. Yang H-C, Islam MM, Jack Li Y-C. Potentiality of deep learning application in healthcare. Comput Methods Programs Biomed 2018;161:A1.

18. Altinier S, Sarti L, Varagnolo M, Zaninotto M, Maggini M, Plebani M. An expert system for the classification of serum protein electrophoresis patterns. Clin Chem Lab Med 2008;46:1458–63.

19. Kratzer MA, Ivandic B, Fateh-Moghadam A. Neuronal network analysis of serum electrophoresis. J Clin Pathol 1992;45:612–5.

20. Ognibene A, Graziani MS, Caldini A, Terreni A, Righetti G, Varagnolo MC, et al. Computer-assisted detection of monoclonal components: results from the multicenter study for the evaluation of CASPER (Computer Assisted Serum Protein Electrophoresis Recognizer) algorithm. Clin Chem Lab Med 2008;46:1183–8.

21. Chen R, Jaye DL, Roback JD, Sherman MA, Smith GH. Automated Serum Protein Electrophoresis Interpretation Using Machine Learning-Based Algorithm for Paraprotein Detection. Am J Clin Pathol 2020;154:S7–8.

22. Chabrun F, Dieu X, Ferre M, Gaillard O, Mery A, Chao de la Barca JM, et al. Achieving Expert-Level Interpretation of Serum Protein Electrophoresis through Deep Learning Driven by Human Reasoning. Clin Chem 2021;67:1406–14.

23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825–30.

24. scikit-learn: machine learning in Python — scikit-learn 1.3.1 documentation [Online]. Available at: https://scikit-learn.org/stable [Accessed 25 Jan 2024].

25. Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. 2012 IEEE Conference on Computer Vision and Pattern Recognition 2012;3642–9.

26. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 2015;115:211–52.
27. Breiman L. Random Forests. Mach Learn 2001;45:5–32.
28. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63:3–42.
29. Freund Y. Boosting a Weak Learning Algorithm by Majority. Inform and Comput 1995;121:256–85.
30. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38:367–78.
31. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
32. Liu J, Osadchy M, Ashton L, Foster M, Solomon CJ, Gibson SJ. Deep convolutional neural networks for Raman spectrum recognition: a unified solution. Analyst 2017;142:4067–74.
33. Chollet, François. Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek. Frechen: MITP-Verlags GmbH & Co. KG; 2018.
34. Keras Documentation [Online]. https://keras.io [Accessed 25 Jan 2024].
35. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
36. Guyon I, Gunn S, Nikravesh M, Zadeh LA. Feature Extraction: Foundations and Applications. Berlin: Springer; 2008.
37. Atkinson KE. An introduction to numerical analysis. New Jersey: John Wiley & Sons; 2008.
38. Graf O. Arbeitsphysiologie. Berlin: Springer; 2013.

39. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems 35 (NeurIPS 2022) 2022;507–20.

40. Hu H, Xu W, Jiang T, Cheng Y, Tao X, Liu W, et al. Expert-Level Immunofixation Electrophoresis Image Recognition based on Explainable and Generalizable Deep Learning. Clin Chem 2023;69:130–9.

41. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019;364:l886.

42. Poon AIF, Sung JJY. Opening the black box of AI-Medicine. J Gastroenterol Hepatol; 2021;36:581–4.